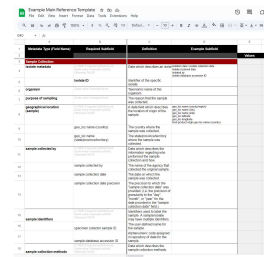


Jun 06, 2025

Contextual Data Collection Form Development & Extraction Protocol (Scoping Review)

DOI

<https://dx.doi.org/10.17504/protocols.io.eq2lywrzevx9/v1>



Rhiannon Cameron¹, Tian Rabbani¹, Miguel Dario Prieto¹, Marcus Yee², Jun Duan¹, Charlotte Barclay¹, Kijin Kim², Chao Chun Liu², Anoosha Sehar¹, Aishwarya Sridhar¹, Arien Fucoy¹, Emma Griffiths¹, William Hsiao³

¹Center for Infectious Disease Genomics and One Health, Simon Fraser University Faculty of Health Sciences, Burnaby, British Columbia, Canada;

²Center for Infectious Disease Genomics and One Health, Simon Fraser University Department of Molecular Biology and Biochemistry, Burnaby, British Columbia, Canada;

³Center for Infectious Disease Genomics and One Health, Simon Fraser University Faculty of Health Sciences & Department of Molecular Biology and Biochemistry, Burnaby, British Columbia, Canada

Rhiannon Cameron: Conception, Design, Templates, Validation, Troubleshooting, User Feedback, Acquisition of Data;

Tian Rabbani: Conception, Design, Troubleshooting, User Feedback, Acquisition of Data;

Miguel Dario Prieto: Troubleshooting, User Feedback, Acquisition of Data

Marcus Yee: Troubleshooting, User Feedback, Acquisition of Data

Jun Duan: Python code, File Organization, Troubleshooting, User Feedback, Acquisition of Data

Charlotte Barclay: Troubleshooting, User Feedback, Acquisition of Data

Kijin Kim: Troubleshooting, User Feedback, Acquisition of Data

Chao Chun Liu: Troubleshooting, User Feedback, Acquisition of Data

Anoosha Sehar: Troubleshooting, User Feedback, Acquisition of Data

Aishwarya Sridhar: Troubleshooting, User Feedback, Acquisition of Data

Arien Fucoy: Acquisition of Data

Emma Griffiths: Practice Contextual Data (Metadata) Extraction Scenario

William Hsiao: Conception, Design



Rhiannon Cameron

Center for Infectious Disease Genomics and One Health

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.eq2lywrzevx9/v1>

External link: <https://github.com/tianrabbani/ScopingReview>

Protocol Citation: Rhiannon Cameron, Tian Rabbani, Miguel Dario Prieto, Marcus Yee, Jun Duan, Charlotte Barclay, Kijin Kim, Chao Chun Liu, Anoosha Sehar, Aishwarya Sridhar, Arien Fucy, Emma Griffiths, William Hsiao 2025. Contextual Data Collection Form Development & Extraction Protocol (Scoping Review). **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.eq2lywrzevx9/v1>

Manuscript citation:

Rabbani, T., Cameron, R., Prieto, M.D., Yee, M., Duan, J., Barclay, C., et al. (TBD) Metadata usage in infectious disease genomic epidemiology applied to outbreak investigation and surveillance settings: a scoping review. <Manuscript undergoing submission, 2025>

License: This is an open access protocol distributed under the terms of the **[Creative Commons Attribution License](#)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: September 16, 2024

Last Modified: June 06, 2025

Protocol Integer ID: 107723

Keywords: metadata, qualitative, data extraction, data collection form, public health, contextual data, contextual metadata, extraction of contextual metadata, contextual metadata field, presence of contextual metadata, frequency counts of contextual metadata field, metadata, contextual data collection form development, research setting, qualitative protocol, scoping review, correlation with research setting, extraction protocol, multiple data extractor, narrative review, data, manuscript, manual review, derived data, study, different study setting, objective, dynamic data collection form



Funders Acknowledgements:

Canadian Institutes of Health Research Canadian Graduate Scholarship for Doctoral Students

Grant ID: 187519

Interoperable and Distributed Data Ecosystem for Agrifood Systems (IDDEAS)

Grant ID: 322DCC

Developing international best practices for Wastewater Surveillance

Grant ID: Investment ID INV-064277, ORS Proposal: 29633

"Building the Canadian Public Health Genomics Contextual Data Harmonization and Sharing Ecosystem"

Grant ID: 2223-HQ-000265

Vocabulary, Thesaurus and Ontology Development for USDA FoodData Central

Grant ID: 58-8040-8-014-F

Abstract

This is a qualitative protocol that helps users develop dynamic data collection form templates for the extraction of contextual metadata by multiple data extractors. Manuscripts and other resources undergo manual review to assess the presence of contextual metadata, that accompanies sequence-derived data, by grouping them in flat or hierarchical fields and observing their correlation with research settings/objectives. The results can enable the user to explore the frequency counts of contextual metadata fields for different study settings/objectives, while also setting the user up for the option of a narrative review of in-text quotes that argue in favour of or against the use/collection of specific types of contextual metadata.

Image Attribution

"Scoping Review Data Collection Template" by [Rhiannon Cameron](#) is in the [Public Domain, CC0](#)

Guidelines

Timings are not listed on this protocol since steps are not time-sensitive and will be dependent on the individual executing this workflow as well as the time it takes to iteratively work through certain steps with collaborators. Please note this process may take weeks or months depending on the number of studies you are extracting from and the amount of specification fields you chose to identify values for.

Materials

- Google Sheets

Safety warnings

- ⚠ There are many data items during data collection that could be easily inferred from other values. For example, when provided the sample "collection device" as "nasopharyngeal swab", one could infer the "collection site host" was the "nasopharynx". We recommend deciding before data extraction begins the degree of inferences that are permitted during data extraction.

You may want to avoid allowing inferences so that you firstly, know which data fields were addressed directly/explicitly by the authors so you can evaluate which data elements were known to be present for addressing different research questions/objectives. Secondly, avoid inadvertent data transformation. For example, a paper states that a sample was collected in "Nashville", so one may infer that the state at which sample collection occurred was "Tennessee" without realizing there is actually another "Nashville" located in "Indiana". Or perhaps a data extractor searches the location of a hospital in a foreign country and believes they have identified the best regional location information, but this country doesn't have English as a first language and more accurate location descriptions may not have been accessible to the curator who is not literate in the native language. Or when given the common name "mink" a data extractor accidentally infers the scientific name for the American mink (*Neovison vison*) rather than the European Mink (*Mustela lutreola*). Experienced curators have practice at avoiding these mistakes, but new data extractors don't have extensive curation training so it may be worth eliminating ambiguity and variability during data collection by avoiding inferences whenever possible.



Before start

This protocol assumes you have already undergone the process of "Quality Appraisal", where careful assessment has been undergone to ensure the documents being used for extraction meet all the necessary criteria for the study. E.g. PRISMA guidelines for a scoping review.

Read over the protocol to make sure this is the right fit for you needs. Consider the "Analysis" section and whether the resulting outputs will be appropriate for your research objectives; will you perform the same analyses or wish to use the outputs in a different way? If you want to use them in another way, does the collection method actually give you data capable of exploring those questions?

EXAMPLE: For your analysis you decide you want to see what kinds of sequencers were listed in the "Metadata Table" results; exploring their presence across studies and how they changed over time. However, the data collection protocol was just looking for "TRUE" values, where the identification of one sequencer is all that was needed to be positive. This means that the data extractors didn't necessarily keep looking for all sequencers mentioned within the study because they could stop after one instance, since as far as they are concerned that is sufficient and efficient. If you want all values and not just the first one, you need to make that clear in the protocol before data extraction begins (or go back and redo the data extraction) and adjust your analysis plan accordingly, otherwise your "sequencer" results will not be comprehensive nor useful.

Consider your objectives and review the "Analysis" section to see if the outputs presented here are an appropriate fit. If not, consider if the protocol can be modified to support your objectives. This protocol was for a scoping review, but we recommend you review qualitative data extraction review practices before proceeding, such as this LibGuides Systematic Review Data Extraction Methodology:

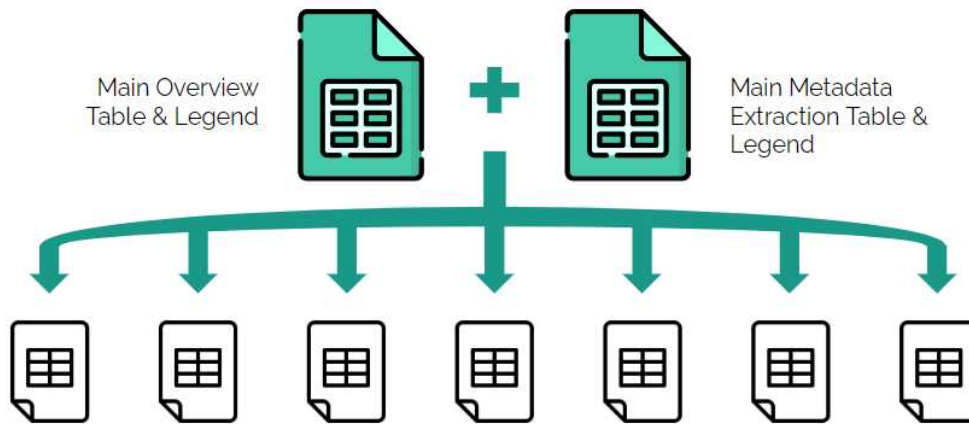
https://ifis.libguides.com/systematic_reviews/data-extraction

Data extraction forms preparation

- 1 You should start by drafting a list of variables of interest with your team. We recommend separating out overarching variables regarding a study (e.g. open-source, study type, etc.) from the contextual metadata value identification. The "**Overview Table**" needs to include the characteristics and key information we want to capture while the "**Metadata Table**" will focus on the contextual metadata "fields" or "values" of interest for your work. If you are not intending on extracting contextual metadata you can ignore the actions regarding it. Discuss which variables are of interest with the team and contextual metadata advisors (if available), and then draft the data collection format based on how they could be used in downstream analysis.
- 2 **Open up a new worksheet** (i.e. collection of spreadsheets) and make the following tabs:
 - **README**
 - **Overview Table**
 - **Overview Legend**
 - **Metadata Table**
 - **Metadata Legend**

Software	
Sheets	NAME
	DEVELOPER
	SOURCE LINK
Google	
https://docs.google.com/spreadsheets	

If there will be more than one data extractor we recommend using "Google Sheets" or another online collaborative spreadsheet tool so that there can be a "**Main/Master/Reference**" version of the template that all other instances of individual worksheets can import from. This ensures that **(A)** when there are changes to any of the worksheets (e.g. clarification in a definition or a variable change) it will be easier to propagate all changes across all active worksheets, and **(B)** data extractors don't accidentally change their reference materials such that they are no longer aligned in their understanding of terms/protocols during data extraction.



Dynamic worksheets. "Main/Master/Reference" versions of the template are imported to individual sheet.

Do not worry about formatting at this point, you are still in the drafting stage. The "**Table**" tabs will eventually turn into the data extraction forms, while the "**Legend**" tabs will be where extraction variables are described. The "**README**" will be information for the data extractor about the data extraction worksheet and how to use it; this should be the opening tab for the worksheet.

3 **Add versioning information** to all tabs.

Put the version somewhere you can easily see when the page is first loaded. A basic approach to versioning is "version x.y" where "x" are changes to the overall structure of the template (e.g., the addition or removal of variable columns/rows) and "y" changes are minor (e.g. changes to a definition or an example that were already present).

You can also get more sophisticated by adding additional versioning levels, e.g. "x.y.z.+" where you assign specific types of changes to each increment - usually keeping the largest scale changes at "x" level. This will be easier to do once you are out of the draft stage and have an initial template that you can conceptualize applying the versioning to.

Since we are still in the drafting stage ($x = 0$), you can keep things simple and **begin with "0.1"**, in the future applying all changes as increments in the "y" spot until you release your pilot version at "1.0".



Note

IMPORTANT: Our approach was to have the **individual tabs have their own versions** because the component parts end up separated out in the future and sometimes changes in one don't affect all the others. However, another possibility would be to version the entire template, e.g. where "x" are considered major changes, "y" are considered template changes", and "z" are considered legend and/or readme changes. Do what is intuitive to you and that you know you can execute consistently.

Note

HIGHLIGHT/ANNOUNCE VERSION CHANGES: It can be helpful to make version changes obvious, perhaps even tracking them in a separate "version tracking" tab (or perhaps in the "README") to help you and data collectors review what has changed between releases.

Draft the "README"

- 4 We recommend that a "horizontal layout" (variables in rows) is used for this spreadsheet to ensure that the text layout is more suitable for a normal "top to bottom" reading experience. Remember you are still in the drafting stage so your descriptions can be simple to start, they will become more comprehensive with relevant information as you go along. To help you get started you can pull from the "EXAMPLE" below.

Include at minimum:

- **Data Extractor Name** - Full name of the data extractor.
- **Date Range of Extraction** - For the start and end date of extraction; you can indicate the format with placeholder text.
- **General Instructions** - General instructions for how users should use the data extraction form.
- **Resource Location** - How and where users can access the resources they will be extracting from.
- **Overview Table** - Description of the table and useful user information.
- **Metadata Table** - Description of the table and useful user information.

Optional recommendations:

- **Formatting/Customization** - Information on the limitations of personal customization of the table.
- **Special Table Variable(s)** - Variables whose protocol require a different approach from the majority.
- **Additional Contact Information** - Contact information for the data extractor and/or project lead(s).



- **Troubleshooting** - Troubleshooting information for situations that are predictable or experienced during extraction.

EXAMPLE

A	B	C
Data Extractor Name:	<First Name> <Last Name>	version 0.1
Date Range of Extraction:	<YYYY-MM-DD> to <YYYY-MM-DD>	
General:	This is your personal worksheet for the scoping review full-text metadata extraction. All headers are imported from a master reference sheet that will be updated as necessary to ensure everyone has the most up-to-date and comprehensive information available. It is recommended you familiarize yourself with and frequently reference the "Legend" tabs, as they are your guide to filling in the tables. All data extraction tables have a aforementioned "Legend" that describes headers, and provides value examples/descriptions of what those fields are for. Tables are designed so that they import reference information from a reference sheet; when the reference sheet is updated the changes appear on individual data collection sheets. This keeps data collectors in sync when it comes to the table "Legends", headers, etc. These imported cells are protected so that only the project lead (<First & Last Name>) or the data extract worksheet manager (<First & Last Name>) can edit them to avoid users from accidentally breaking a formula. For data extraction examples, we recommend you review the information provided on your personal scoping review full-text metadata extraction pilot worksheet. All data extraction is to be performed manually as described in the protocol.	
Manuscript Location:	You will find the link to the manuscript you've been assigned under the "Nextcloud (Data Storage Location)" column. Make sure you double check that the title/date in the file matches the information in the overview table. If the file is missing, corrupted, or doesn't appear to be a	

A	B	C
	manuscript, login to Covidence and search for your article using the "Covidence #" with the "#" sign in front.	
Overview Table:	The "Overview Table" is a summary table that provides data extractors with relevant manuscript information and overarching variables of interest the scoping review team and metadata advisors have advised the capture of. Your copy of the table will only contain the manuscripts you have been assigned for data extraction.	
Metadata Table:	The "Metadata Table" is where more granular metadata evidence is captured to indicate the presence or absence of a metadata type/field. "Metadata Type (Field Name)"s are deemed "TRUE" if a required subfield was identified, or in the case where an example subfield value was found, indicated as TRUE by the inclusion of the quoted "value" which supports the conclusion. Otherwise, indicate "FALSE" to confirm you tried but could not find evidence of a data item.	
Formatting/Customization & Troubleshooting:	To make the data extraction worksheet more comfortable for yourself, you are welcome to customize the formatting of cells provided you are customizing the entire cell and not just a selection of text. This is because trying to format a smaller range of text will break the import function and result in a "REF#" error. If you receive this error you can resolve it by undoing the formatting you have applied. You should also not rearrange and columns or rows so as to ensure appropriate data alignment when your tables are merged into the masters.	

Example "README" worksheet tab.

Resource manager data extraction

5 **Extract/export metadata on the data sources from whatever resource manager you are using.**

Keep a copy of files for provenance and backup. We encourage you to include RAW and processed data as part of a data package to publish with your final work.

Software

GitHub

NAME

GitHub Inc.

DEVELOPER

<https://github.com/>

SOURCE LINK

GitHub is a great tool for backing up and sharing your data (you can always keep the repository private until you are ready to go public). That being said do **de-identify the data** before release or uploading to GitHub. Even if you overwrite an upload, GitHub will keep the information in its commit history and people may use that to find identifying information. It is easier to upload de-identified data to begin with even if in the end you decide to switch to including identifying information.

5.1 Resource Manager Data Extraction: COVIDENCE

Source: <https://support.covidence.org/help/exporting-your-extracted-data> (accessed September 2024)

Software

Covidence

NAME

<https://www.covidence.org/>

SOURCE LINK

Login to COVIDENCE.

1. Go to the "**Summary Page**" of your review.
2. Select the "**Export**" tool.
3. Under "Type" select "**Data Extraction**".
4. Under "Data" select "**Inclusion**".
5. Under "Format" select "**Single sheet (CSV)**".
6. Select "**Prepare file**".
7. Your file will be ready when the "**Download**" button appears.
8. Select "**Download**" and save your file.

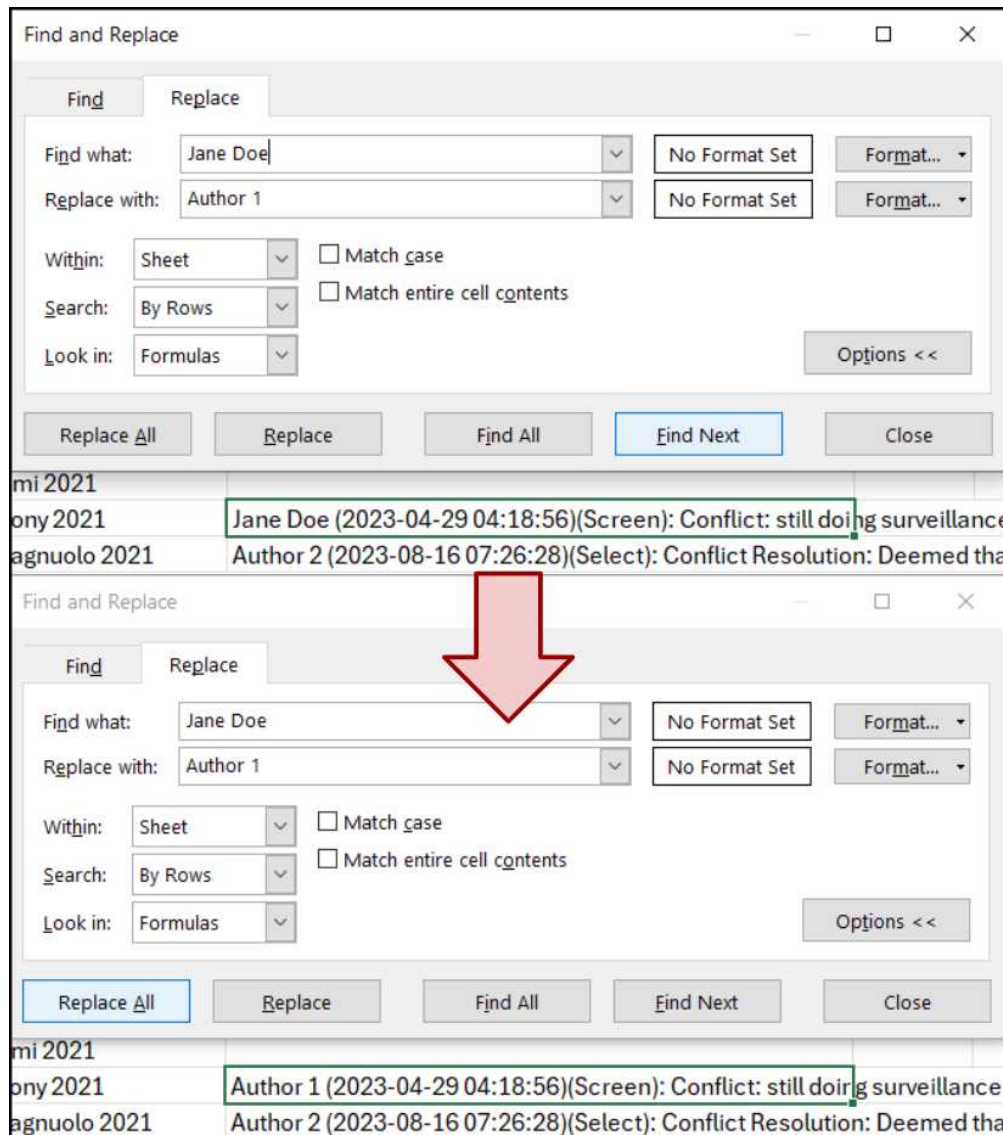
**Note**

Keep a copy of this RAW file for provenance and backup, although you may want to proceed with Step 4.2 before doing so.

6 De-identify COVIDENCE Export (OPTIONAL)

Open a spreadsheet application and use "Find and Replace" to find the reviewer names. It may also be worth searching for first names on their own and nicknames in case they were referenced by another reviewer within a comment; be careful to check these cases individually so you don't accidentally overwrite authors with the same name listed for the manuscripts that were reviewed.

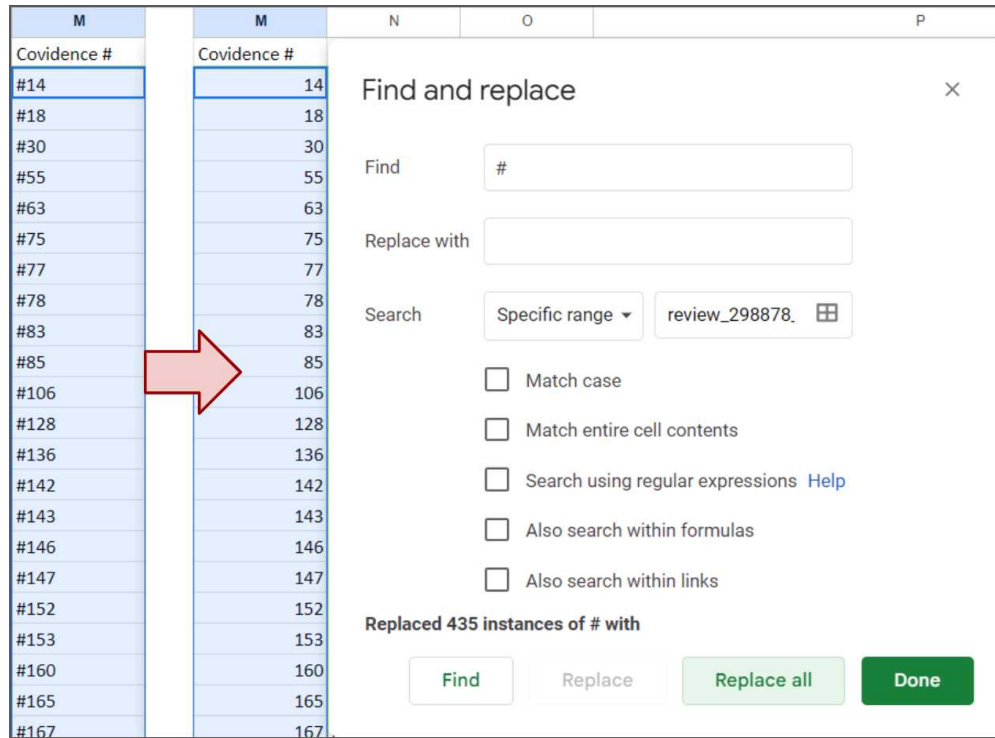
You can make a key and assign author numbers that you can reference later in case you need to reverse this in the future and no longer have the version that hasn't been de-identified.



De-identification using "Replace"/"Find and Replace" in Microsoft Excel.

7 Clean-up COVIDENCE # (OPTIONAL)

We recommend removing the "#" from the "Covidence #" data now, as otherwise it can get in the way of data processing since it makes the cell behave as a non-number.



Removal of the "#" number sign using "Find and replace" in Google Sheets. The "Search" space has been narrowed to the "Specific range" of the relevant column contents.

Draft the "Overview Legend" tab

8 Resource Manager Data Import

1. Go to Sheets and open a blank spreadsheet.
2. Go to **"File"** in the ribbon and *select "Import"*.
3. *Navigate* to your file and *upload* it.

This can be uploaded to your current worksheet or a new one. We will be using some of the variables (i.e. columns) to jumpstart the "Overview" tabs.

9 Getting Started

The **"Legend"** tables are a useful place to initially plan your extraction tables because this will ultimately be your reference guide for your variables. You can capture discussion notes and integrate all relevant information as needed. We recommend that a "horizontal layout" is used for this spreadsheet to ensure that the text layout is more suitable for a normal reading experience.

Give your table the headings:

- **Header/Field** - A column for the name of the headers/fields that will appear for the columns of the data extraction "Overview Table".
- **Description** - A description of the "Header/Field".
- **Values** - The values data collectors can input (e.g., TRUE or FALSE, free-text, etc.), including missing data values ("Not Applicable", "Not Collected", etc. *see "Note" for more information*).
- **Value Description** - A description of the input values for data collectors to help them understand and choose the appropriate input for their situation.
- **Data Type** - The data type of input values (e.g., string, char, int, float, enum, etc. - *see the "Note" below for more information*). This is important for planning your downstream data use but can be removed from the finished version if deemed redundant or confusing for data collectors.

EXAMPLE

	A	B	C	D	E	F
	Header	Description	Data Type	Values	Value Description	Version 0.1

Initial headers for the "Overview Legend" table.

Note

More information on **"Data Type"**: <https://amplitude.com/blog/data-types>
 More information on **"Missing Value Reporting Terms"**:
<https://www.insdc.org/technical-specifications/missing-value-reporting/>

9.1 Now it is time to start **listing the columns/headers we want to eventually see in the "Overview Table"**. We recommend including, at minimum, the following headers/fields (or whatever can be similarly found) from your extracted reference list:

- **Study/Covidence #** - A unique code assigned to the manuscript. If your references export does not have a code like this you can number them once the studies are listed in your "Overview Table" (e.g. using Sheet's "**autofill to complete a series**" function).
- **Study** - A shorthand label for the manuscript that includes the first author's last name and publication year (e.g. "Baier 2022").
- **Title** - The title of the manuscript.
- **Authors** - The manuscript author list.
- **DOI** - A persistent digital object identifier (DOI) that can direct you to the journal publication.

Additional recommended fields:

- **Data Storage Location** - A URL directly to the manuscript or to a folder containing the manuscript. We used a cloud service (e.g. **Nextcloud**) to organize the manuscripts into folders labelled with assigned Study #. Steps on how to prepare these are under "Preparing files for contextual metadata extraction".

Optional:

- **Assignee** - The assigned data extractor when working in a group.
- **Variables of Interest** - a header to help separate the layout of your variables of interest from other data elements.

EXAMPLE

	A	B	C	D	E	F
	Header	Description	Values	Value Description	Data Type	Version 0.1
	Assignee					
	Study/Covidence #					
	Study					
	Title					

	A	B	C	D	E	F
	Authors					
	DOI					
	Data Storage Location					
	Variables of Interest					

Version 0.1. The beginning of forming the "Overview Legend" table.

- 9.2 **Add descriptions to these general "Headers/Fields".** You can also add "Date Type" and "Values" if you please. We considered this optional for these initial header/fields since we weren't going to use them for downstream analysis and the data extractors would already have this information pre-filled by the collection form manager.

EXAMPLE

	A	B	C	D	E	E
	Header	Description	Values	Value Description	Data Type	Version 0.1
	Assignee	The team member assigned to extra metadata from the associated article.				
	Study/Covidence #	The Covidence document reference number (#).				
	Study	Shorthand study label indicating <1st author> .				
	Title	Manuscript/Document title.				
	Authors	Manuscript/Document author list.				
	DOI	DOI reference ID or PURL.				

	A	B	C	D	E	E
	Data Storage Location	URL link to the data storage folder on the Centre for Infectious Disease Genomics and One Health (CIDGOH) Nextcloud server. This is where manuscripts can be found/uploaded, and where to upload additional data such as supplementary files and sequence data (when applicable).				
	Variables of Interest	Overarching variables we've chosen to extract for analysis.				

Version 0.1. Adding descriptions to the current selection of headers.

- 10 **Developing your "Variables of Interest" (an iterative process).** Brainstorm variables of interest (i.e., characteristics and key information) that you want to capture to explore your research question and objectives. The way you format your values will influence downstream data analysis and the structure of the collection forms to facilitate collection and analysis. This will be an iterative development process, only to be truly finalized after you have completed the controlled pilot data extraction.

EXAMPLE

	A	B	C	D	E	E
	Header	Description	Values	Value Descript ion	Data Type	Version 0.1
	<i>[----- removed for brevity-----]</i>					
	Variables of Interest	Overarching variables we've chosen to extract for analysis.				



	A	B	C	D	E	E
	Sequence Data Accessible		TRUE/FALSE		boolean	
	Metadata Table Accessible	- Are we thinking supplementary tables or in-text too?	TRUE/FALSE		boolean	
	Open Access		TRUE/FALSE		boolean	
	Organism	List the name(s) of the infectious pathogens under study.			string	
	Organism Type		Virus, Bacteria		enum (use dropdown list)	

Version 0.1. First attempt at drafting metadata variables

Note

CHECK: Are people allowed to put multiple values upon data entry or just one (e.g. for non TRUE/FALSE cases)?

Note

CHECK: Are any of these perhaps more appropriate for the "Metadata" tables? If so, move or copy into the "Metadata Legend" for consideration.

10.1

Meet with collaborators and data extractors. Meet with your team and metadata advisors to continue brainstorming variables of interest that support your research question, objectives, and facilitate downstream analysis. Capture instructions that are useful for data extractors to know because there should be specific guidance for how a variable is to be collected.

Also decide what separator you will be using and include it in the header of any fields where more than one value is allowed. We recommend you be consistent by using one separator across all data entry; the "|" (pipe) is our top choice as it avoids the issues of using the "," (comma) and extracting to CSV, and is unlikely to appear in "in-text quotes" if you are capturing that kind of data.

EXAMPLE

A	B	C	D	E	E
Header	Description	Values	Value Description	Data Type	Version 0.2
[----- removed for brevity----- -]					
Variables of Interest	Overarching variables we've chosen to extract for analysis.				
Sequence Data Accessible	We want this to help us evaluate sequence data sharing and accessibility. Need to double check it is actually shared and not just claimed to be shared.	Database Name, Not Applicable		string	
Sequence Data Recorded	Discusses saving sequence information on the lab server in space dedicated for this purpose. Store in a folder labelled with the covidence number (requested by lab supervisor).			boolean ?	
Metadata Table Accessible	- Are we thinking supplementary tables or in-text too?	TRUE/FALSE		boolean	
Open Access	If the document was was available via open access (i.e. not behind a paywall). IF not indicated on the journal/document provided, this can be checked by (a) checking the DOI link in incognito mode on your browser,	TRUE/FALSE		boolean	

	A	B	C	D	E	E
		and (b) searching the university library catalogue to see if the journal is listed as OPEN (and where)...				
	Organism Type	Capture this so we can see what the overview of organisms types are.	Virus, Bacteria, Fungi, Protozoa		enum	
	Organism	- Supervisor noted this would be of value to the lab for people interested in doing a similar but more pathogen specific version of this scoping review / contextual metadata collection. - Don't need to list at the variant level, probably not even the subspecies level. NCBITaxon could be used for rounding up to more general classes.	List the name(s) of the infectious pathogens under study.		string	

Version 0.2. Second attempt at drafting variables in discussion with other team/lab members. Discussion notes are interwoven with definitions (when present) in the descriptions column.

Note

CHECK: Are people allowed to put multiple values upon data entry or just one (e.g. for non TRUE/FALSE cases)?

Note

CHECK: Are any of these perhaps more appropriate for the "Metadata" tables? If so, move or copy into the "Metadata Legend" for consideration.

- 10.2 **Fill in the details and separate values out onto individual lines, clearly defining each one.** Clean-up the description and move value specific information to the appropriate "Value Description". When pulling descriptions/definitions from other sources be sure to reference them (e.g. include a URL to your source). If you are allowing more than one

data value in a field be sure to include the separator information in the description and in the table header when you make the "Overview Table".

EXAMPLE

A	B	C	D	E	E
Header	Description	Values	Value Description	Data Type	Version 0.5
[----- removed for brevity-----]					
Variables of Interest	Overarching variables we've chosen to extract for analysis.				
Type	If the document type is surveillance, outbreak investigation, or both. Do you best to pick a primary study type of either "surveillance" or "outbreak investigation" (e.g. a surveillance study that references historical outbreaks would be "surveillance" but not "outbreak investigation", and vice versa for an "outbreak investigation" that happens to reference "surveillance data"). Only use "Both" in	Surveillance	We are working with the understanding that surveillance is the ongoing systematic collection, analysis, and interpretation of health data that are essential to the planning, implementation, and evaluation of public health practice. This means a surveillance study can be deduced based on if isolates/data are collected over a span of time (such as with contact tracing); otherwise err on the side of not surveillance if you're not 100% sure whether the study is conducting surveillance (i.e. if the study is using surveillance isolates but not conducting actual surveillance, etc.). If a study describes itself as surveillance then we defer to the study. (source:	enum	



A	B	C	D	E	E
	cases of extreme certainty or uncertainty. Note: Please refer to the "Glossary" for more study type descriptions and other agreed upon terminology.		https://www.sciencedirect.com/science/article/abs/pii/B9780127640518500421)		
		Outbreak Investigation	A study is an outbreak investigation if the authors describe investigating an outbreak. Note that a cluster of cases is not necessarily and outbreak. Generally, the aim of outbreak epidemiology is to study an epidemic in order to gain control over it and to prevent further spread of the disease. Generally outbreak means a 'sudden occurrence,' while in the epidemiological sense an outbreak is defined as a sudden increase in the disease frequency, related to time, place, and observed population. (source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7187955/)	enum	
		Both	A surveillance context in which activities are specifically triggered by the detection of [a disease] outbreak. (source: http://purl.obolibrary.org/obo/HSO_0000371)	enum	
Sequence Data Accessible (Database Name)	If the sequence data is publically available and accessible through information provided by the article, indicate which database it is located at.	[Database Name]	The name of the database where the sequence data was located. Note: Captured as a string but can be converted (>) to a "TRUE" boolean for analysis.	string > boolean	
		FALSE	Database information could not be identified.	boolean	



A	B	C	D	E	E
Sequence Data Saved (TRUE/FALSE)	Please save sequence information on the lab server space dedicated for this purpose. Store in a folder labelled with the manuscript/document Covidence number. ONLY SAVE THE DATA if stored in a non-mainstream repository. Mainstream repositories are deemed more likely to be accessible for future research and thus not necessary to retrieve data from. You may note down the and/or or leave it to researchers to identify this information from the manuscript.	TRUE	Sequence data or information about it has been stored on the CIDGOH server.	boolean	
		FALSE	Sequence data or information about it has not been stored on the CIDGOH server.	boolean	
Open Access (TRUE/FALSE)	If the document was available via open access (i.e. not behind a paywall). If not indicated on the journal/document provided, this can be checked by (a) checking the DOI link in incognito mode on your browser, and (b) searching	TRUE	An open access version of the manuscript/document was identified.	boolean	



	A	B	C	D	E	E
		the institutional Library catalogue to see if the journal is listed as OPEN (and where). It is best to perform both checks as just because the DOI link provided by Covidence isn't open, doesn't mean it isn't openly accessible at another journal.	FALSE	An open access version of the manuscript/document was not identified.	boolean	
	Organism Type (Virus Bacteria Fungi Protozoa)	Infection pathogens that undergo whole genome or near complete genome sequencing. Not including worms or prions. Some studies include more than one type, in which case indicate all applicable organisms, separated by a " " (vertical bar/pipe).	Virus	An infectious agent which consists of two parts, genetic material and a protein coat. These organisms lack independent metabolism, and they must infect the cells of other types of organisms to reproduce. Most viruses are capable of passing through fine filters that retain bacteria, and are not visible through a light microscope. (source: http://purl.obolibrary.org/obo/NCIT_C14283)	enum	
			Bacteria	Unicellular, prokaryotic organisms that reproduce by cell division and usually have cell walls; can be shaped like spheres, rods or spirals and can be found in virtually any environment. (source: http://purl.obolibrary.org/obo/NCIT_C14187)	enum	
			Fungi	A kingdom of eukaryotic, heterotrophic organisms that live as saprobes or parasites,	enum	



A	B	C	D	E	E
			including mushrooms, yeasts, smuts, molds, etc. They reproduce either sexually or asexually, and have life cycles that range from simple to complex. Filamentous fungi refer to those that grow as multicellular colonies (mushrooms and molds). (source: http://purl.obolibrary.org/obo/NCIT_C14209)		
		Protozoa	Unicellular heterotrophic eukaryote in the kingdom protista. (source: http://purl.obolibrary.org/obo/NCIT_C77916)	enum	
Organism (separator =)	List the name(s) of the infectious pathogens studied. Indicate all applicable organisms, separated by a " " (vertical bar/pipe). Note: You do not need to list at the variant level.	[Organism Name]	Latin scientific name for the organism of study.	string	
Metadata Accessible	If a metadata table is available or not.				
In-Text (TRUE/FALSE)	Metadata available from the main body text of the manuscript/document. This value does not indicate the presence of a metadata table, but is still a variable of interest.	TRUE	Metadata available within text.	boolean	
		FALSE	Metadata not available within text.	boolean	
Manuscript Embedded	Metadata available as	Summary	Metadata available within embedded table	string >	

	A	B	C	D	E	E
	Table (Summary Table/Line List/FALSE, separator =)	an embedded table within the manuscript/document. The presence of both can be indicated by listing them using the " " separator (vertical bar/pipe).	Table	as a aggregated data (e.g. listing out counts/date for a specific demographic and associated symptoms). Note: Captured as a string but can be converted (>) to a "TRUE" boolean for analysis.	boolean	
			Line List	Metadata available within embedded table as a individual data points (i.e. individual subjects and their metadata values). Note: Captured as a string but can be converted (>) to a "TRUE" boolean for analysis.	string > boolean	
			FALSE	Metadata not available within embedded table.	boolean	
	[---see manuscript dataset for more examples---]					

Version 0.5. A preview of what the "Overview Legend" may look like after multiple iterations.

Draft the "Overview Table" tab

11 Create the "Overview Table" collection template in the "Overview Table" tab.

- The headers should be the same as those listed in the "Overview Legend", but flipped so that they are in a "vertical table" format (i.e., variables along the top row). This can be easily done by copying the "Header/Field" column from the "Overview Legend" and then pasting transposed in the "Overview Table".
- Be sure to include your separator value in the applicable headers, you can also include the input values.

- For enumerated (enum) lists you can make a picklist/dropdown menu (Note: this was not done for "Organism Type" in the example below because this field allowed for the input of one or more items; the Sheets dropdown is only single select so use a pipe "|" to create a separated list).
- Don't forget your version information.
- OPTIONAL: "Freeze" the header rows as default so that they remain in view as users enter data.

EXAMPLE

[illegible]

Empty "Overview Table" formatted for data entry.

11.1 Add the data relevant data extracted from "Resource manager data extraction".

EXAMPLE

Variables of Interest											
Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type	Sequence Data Accessible (Database Name/FASE)	Sequence Data Saved (TRUE/FALSE)	Open Access (TRUE/FALSE)	Organism Type (Virus Bacteria Fungi Protozoa)
	14	Lin 2021	Genomic monito	Lin JW; Tang C; V	10.1016/j.chom.		▼				
	18	Snitkin 2012	Tracking a hospiti	Snitkin ES; Zelazi	10.1126/scitrans		▼				
	30	Lachmann 2022	Invasive listerios	Lachmann R; Hal	10.1080/222217		▼				
	55	Avetyan 2022	Molecular Anal	Avetyan D; Hako	10.3390/v14051		▼				
	63	RusA' AjkovA' Z	Systematic Geno	RusA' AjkovA' D;	10.3390/v14112		▼				
	75	Mustapha 2022	Genomic Diversi	Mustapha MM;	10.1128/msyste		▼				
	77	Enger 2022	A tale of two STs	Enger H; Larssen	10.1007/s10096		▼				
	78	Ntoumi 2021	Genomic surveill	Ntoumi F; Mfout	10.1016/j.ijid.20		▼				
	83	Brehony 2021	Establishment of	Brehony C; Lanig	10.1111/zph.12f		▼				
	85	Gravagnuolo 20	High throughput	Gravagnuolo AM	10.1016/j.ebiom		▼				
	106	SaavedraCervere	Fine-grain popul	Saavedra Cerver	10.1099/mgen.0		▼				
	128	Sy 2022	Plasmodium falc	Sy M; Deme AB;	10.1038/s41598		▼				
	136	Simon 2019	Whole Genome	Simon B; Pichon	10.3390/v11020		▼				
	142	Kong 2016	Whole-Genome	Kong Z; Zhao P;	110.1371/journal		▼				
	143	Krizova 2019	Genomic surveill	Krizova P; Honsk	10.1371/journal		▼				
	146	Enkelmann 2020	Disentangling ou	Enkelmann J; vor	10.1017/S09502		▼				
	147	Javaid 2021	Real-Time Invest	Javaid W; Ehni J;	10.1093/cid/cia		▼				
	152	Dickinson 2019	Whole-genome	Dickinson H; Rea	10.1016/j.jhin.2f		▼				
	153	Sherry 2019	Genomics for M	Sherry NI; Lane	10.1128/JCM.00		▼				
	160	Baier 2022	Target capture si	Baier C; Huang J;	10.1186/s13756		▼				
	165	Houlihan 2018	Use of Whole-G	Houlihan CF; Fra	10.1093/infdis/ji		▼				
	167	Kaku 2022	Changing molecu	Kaku N; Sasaki D	10.1093/jac/dka		▼				
	174	Ugolotti 2018	Genomic charac	Ugolotti E; Di M	10.1016/j.jhin.2f		▼				
	181	CremersAJH 202	Surveillance-eml	Cremers AJH; Co	10.1038/s41598		▼				
	185	Snell 2022	Combined epide	Snell LB; Fisher	10.1016/j.cmi.2f		▼				
	202	Fitzpatrick 2016	Utility of Whole	Fitzpatrick MA;	10.1128/JCM.01		▼				
	203	Kossow 2019	Whole genome	Kossow A; Kamp	10.1016/j.jhin.2f		▼				

The "Overview table" with Covidence manuscript information.

- 11.2 **Add your assignees (if applicable).** Keep in mind that assignment is randomized and each data extractor has received the agreed upon number of extraction assignments. The "Overview" will eventually be subdivided/filtered so that data extractors only get their own list.

There are many ways you can do this. One approach is walked through as an example below.

EXAMPLE

	A	B	C	D	E	F
1	Extractor 1	Extractor 2	Extractor 3			Assignments
2					Extractor 1	6
3					Extractor 2	6
4					Extractor 3	3
5						
6						

Start a temporary sheet for generating your extractor list.



	A	B	C	D	E	F
1	Extractor 1	Extractor 2	Extractor 3			Assignments
2	Extractor 1				Extractor 1	6
3	Extractor 1				Extractor 2	6
4	Extractor 1				Extractor 3	3
5	Extractor 1					
6	Extractor 1					
7						

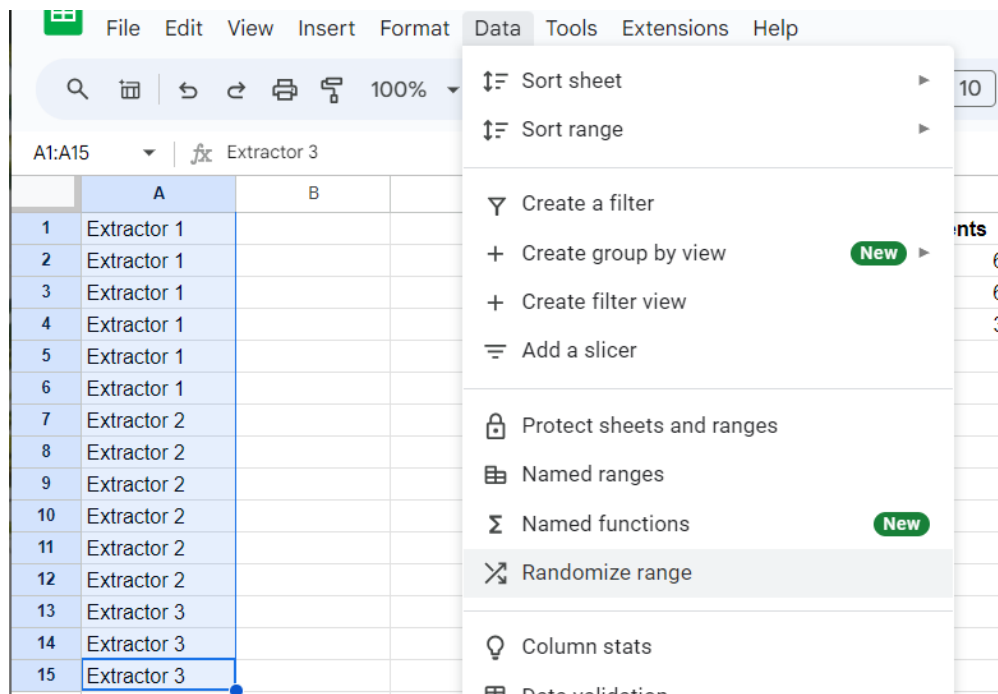
Click and drag "Extractor name" to the row count that matches the assignment count.

	A	B	C	D	E	F
1	Extractor 1	Extractor 2	Extractor 3			Assignments
2	Extractor 1	Extractor 2	Extractor 3		Extractor 1	6
3	Extractor 1	Extractor 2	Extractor 3		Extractor 2	6
4	Extractor 1	Extractor 2			Extractor 3	3
5	Extractor 1	Extractor 2				
6	Extractor 1	Extractor 2				
7						

Repeat for the remaining data extractors. Make sure the counts add up as expected.

	A	B	C	D	E	F
1	Extractor 1					Assignments
2	Extractor 1				Extractor 1	6
3	Extractor 1				Extractor 2	6
4	Extractor 1				Extractor 3	3
5	Extractor 1					
6	Extractor 1					
7	Extractor 2					
8	Extractor 2					
9	Extractor 2					
10	Extractor 2					
11	Extractor 2					
12	Extractor 2					
13	Extractor 3					
14	Extractor 3					
15	Extractor 3					

Combine into one column.



Select column and then "Data" > "Randomize range".

	A	B	C	D	E	F
1	Extractor 2					Assignments
2	Extractor 2				Extractor 1	6
3	Extractor 2				Extractor 2	6
4	Extractor 3				Extractor 3	3
5	Extractor 1					
6	Extractor 3					
7	Extractor 1					
8	Extractor 1					
9	Extractor 1					
10	Extractor 2					
11	Extractor 2					
12	Extractor 1					
13	Extractor 2					
14	Extractor 3					
15	Extractor 1					

Randomized range results.



							Variables of Interest
Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type
Extractor 2	14	Lin 2021	Genomic monitoring	Lin JW; Tang C; V	10.1016/j.chom.		
Extractor 2	18	Snitkin 2012	Tracking a hospital	Snitkin ES; Zelazny	10.1126/scitrans		
Extractor 2	30	Lachmann 2022	Invasive listeriosis	Lachmann R; Halperin	10.1080/222217		
Extractor 3	55	Avetyan 2022	Molecular Analysis	Avetyan D; Hakoian	10.3390/v14051		
Extractor 1	63	Rusakova 2022	Systematic Genomic	Rusakova A; D; H	10.3390/v14112		
Extractor 3	75	Mustapha 2022	Genomic Diversity	Mustapha MM; H	10.1128/msystems		
Extractor 1	77	Enger 2022	A tale of two STs	Enger H; Larssen	10.1007/s10096		
Extractor 1	78	Ntoumi 2021	Genomic surveillance	Ntoumi F; Mfouti	10.1016/j.ijid.20		
Extractor 1	83	Brehony 2021	Establishment of	Brehony C; Lanig	10.1111/zph.128		
Extractor 2	85	Gravagnuolo 2021	High throughput	Gravagnuolo AM	10.1016/j.ebiom		
Extractor 2	106	SaavedraCervera	Fine-grain population	Saavedra Cervera	10.1099/mgen.0		
Extractor 1	128	Sy 2022	Plasmodium falciparum	Sy M; Deme AB	10.1038/s41598		
Extractor 2	136	Simon 2019	Whole Genome	Simon B; Pichon	10.3390/v11020		
Extractor 3	142	Kong 2016	Whole-Genome	Kong Z; Zhao P; Li	10.1371/journal.		
Extractor 1	143	Krizova 2019	Genomic surveillance	Krizova P; Honskova	10.1371/journal.		

Copy & paste into the "Overview Table".

The screenshot shows the Microsoft Excel interface. The 'Data' menu is open, and 'Sort sheet' is selected. The 'Sort sheet' submenu is also open, showing 'Sort sheet by column A (A to Z)' and 'Sort sheet by column A (Z to A)'. The 'Sort sheet by column A (A to Z)' option is selected. The main table is visible in the background, showing the same data as the first table, but with the 'Assignee' column highlighted in blue. The 'Data' menu is open, and the 'Sort sheet' option is selected. The 'Sort sheet' submenu is also open, showing 'Sort sheet by column A (A to Z)' and 'Sort sheet by column A (Z to A)'. The 'Sort sheet by column A (A to Z)' option is selected. The main table is visible in the background, showing the same data as the first table, but with the 'Assignee' column highlighted in blue.

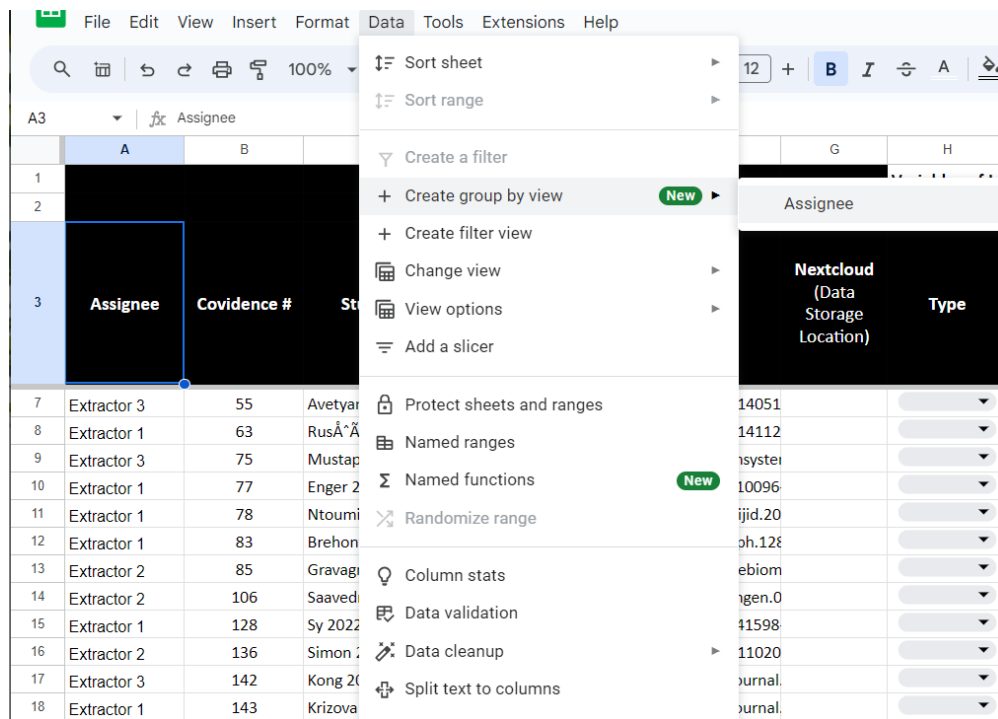
Select column and then "Data" > "Sort Sheet" > "Sort sheet by column A (A to Z)" to group the results by "Assignee". You may need to temporarily remove some rows to make the variable in the first row.

							Variables of Int
Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type
Extractor 1	63	RusÅ~ÅikovÅ 2C	Systematic Geno	RusÅ~ÅikovÅ D;	10.3390/v14112		▼
Extractor 1	77	Enger 2022	A tale of two STs	Enger H; Larssen	10.1007/s10096		▼
Extractor 1	78	Ntoumi 2021	Genomic surveill	Ntoumi F; Mfout	10.1016/j.ijid.20		▼
Extractor 1	83	Brehony 2021	Establishment of	Brehony C; Lanig	10.1111/zph.128		▼
Extractor 1	128	Sy 2022	Plasmodium falc	Sy M; Deme AB;	10.1038/s41598		▼
Extractor 1	143	Krizova 2019	Genomic surveill	Krizova P; Honsk	10.1371/journal.		▼
Extractor 2	14	Lin 2021	Genomic monito	Lin JW; Tang C; V	10.1016/j.chom.		▼
Extractor 2	18	Snitkin 2012	Tracking a hospit	Snitkin ES; Zelaz	10.1126/scitrans		▼
Extractor 2	30	Lachmann 2022	Invasive listerios	Lachmann R; Hal	10.1080/222217		▼
Extractor 2	85	Gravagnuolo 202	High throughput	Gravagnuolo AM	10.1016/j.ebiom		▼
Extractor 2	106	SaavedraCervere	Fine-grain popul	Saavedra Cerver	10.1099/mgen.0		▼
Extractor 2	136	Simon 2019	Whole Genome	Simon B; Pichon	10.3390/v11020		▼
Extractor 3	55	Avetyan 2022	Molecular Analy	Avetyan D; Hako	10.3390/v14051		▼
Extractor 3	75	Mustapha 2022	Genomic Diversi	Mustapha MM; t	10.1128/msyste		▼
Extractor 3	142	Kong 2016	Whole-Genome	Kong Z; Zhao P; l	10.1371/journal.		▼

Results.

Note: You can always switch back to sorting by "Covidence #" by repeating the previous step on the "Covidence #" column.

Alternatively: Instead of sorting by "Assignee" you can experiment with using Sheets' "group by view" feature. However, there will not be additional guidance on using this approach as the feature was not present during the original protocol implementation.



Select column and then "Data" > "Create group by view" > "column A" to group the results by "Assignee".

Assignee	Covidence	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type	Sequence Data Accessible (Database Name/FALSE)	Sequence Data Saved (TRUE/FALSE)	Open Access (TRUE/FALSE)	Organism Type (Virus/Bacterial/Fungi/Protozoa)
Assignee: Extractor 1											
Extractor 1	63	Rusakov 2022	Systematic Gen	Rusakov A, D	10.3390/v14111						
Extractor 1	77	Enger 2022	A tale of two ST	Enger H; Larssen	10.1007/s10091						
Extractor 1	78	Ntoumi 2021	Genomic surveil	Ntoumi F; Mfour	10.1016/j.ijid.20						
Extractor 1	83	Brehony 2021	Establishment c	Brehony C; Lani	10.1111/zph.12						
Extractor 1	128	Sy 2022	Plasmodium fal	Sy M; Deme AB	10.1038/s41591						
Extractor 1	143	Krizova 2019	Genomic surveil	Krizova P; Hons	10.1371/journal						
Assignee: Extractor 2											
Extractor 2	14	Lin 2021	Genomic monitr	Lin JW; Tang C	10.1016/j.chom						
Extractor 2	18	Snitkin 2012	Tracking a hosp	Snitkin ES; Zela	10.1126/scitrans						
Extractor 2	30	Lachmann 2022	Invasive listerio	Lachmann R; H	10.1080/22221						
Extractor 2	85	Gravagnuolo 20	High throughpu	Gravagnuolo AN	10.1016/j.ebion						
Extractor 2	106	SaavedraCerver	Fine-grain popu	Saavedra Cerve	10.1099/mgen.1						
Extractor 2	136	Simon 2019	Whole Genome	Simon B; Pichor	10.3390/v11021						
Assignee: Extractor 3											
Extractor 3	55	Avetyan 2022	Molecular Anal	Avetyan D; Hakk	10.3390/v1405						
Extractor 3	75	Mustapha 2022	Genomic Divers	Mustapha MM	10.1128/msyst						
Extractor 3	142	Kong 2016	Whole-Genome	Kong Z; Zhao P	10.1371/journal						
Assignee: (Blanks)											
	146	Enkelmann 2022	Disentangling o	Enkelmann J; vc	10.1017/S0950						

Results.

Note

CHECK: Make sure there are no "unassigned" studies, or that there is at least a plan for them to be assigned later.

- 11.3 **Add the relevant data extracted from "Preparing files for contextual metadata extraction".** The "Preparing files for contextual metadata extraction" steps can be done in parallel to building these tables. You can add the "Data Storage Location" information at any point onward before full data extraction, although it may be helpful to at least have a subset ready for your data extraction pilot so the data extractors can practice what will most closely resemble the final methodology.

EXAMPLE

Variables of Interest											
Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type	Sequence Data Accessible (Database Name/FASE)	Sequence Data Saved (TRUE/FASE)	Open Access (TRUE/FASE)	Organism Type (Virus Bacteria Fungi Protozoa)
	14	Lin 2021	Genomic monito	Lin JW; Tang C; V	10.1016/j.chom. https://cloud.c		▼				
	18	Snitkin 2012	Tracking a hospit	Snitkin ES; Zelaz	10.1126/scitrans https://cloud.c		▼				
	30	Lachmann 2022	Invasive listerio	Lachmann R; Hal	10.1080/222217 https://cloud.c		▼				
	55	Avetyan 2022	Molecular Anal	Avetyan D; Hako	10.3390/v14051 https://cloud.c		▼				
	63	RusA*ÄikováÄi Z	Systematic Geno	RusA*ÄikováÄi D	10.3390/v14112 https://cloud.c		▼				
	75	Mustapha 2022	Genomic Diversi	Mustapha MM; ;	10.1128/msyste https://cloud.c		▼				
	77	Enger 2022	A tale of two STs	Enger H; Larssen	10.1007/s10096 https://cloud.c		▼				
	78	Ntoumi 2021	Genomic surveill	Ntoumi F; Mfout	10.1016/j.jiid.20 https://cloud.c		▼				
	83	Brehony 2021	Establishment of	Brehony C; Lanig	10.1111/zph.124 https://cloud.c		▼				
	85	Gravagnuolo 20	High throughput	Gravagnuolo AM	10.1016/j.jebiom https://cloud.c		▼				
	106	SaavedraCervere	Fine-grain popul	Saavedra Cerven	10.1099/mgen.0 https://cloud.c		▼				
	128	Sy 2022	Plasmodium falc	Sy M; Deme AB;	10.1038/s41598 https://cloud.c		▼				
	136	Simon 2019	Whole Genome	Simon B; Pichon	10.3390/v11020 https://cloud.c		▼				
	142	Kong 2016	Whole-Genome	Kong Z; Zhao P;	10.1371/journal https://cloud.c		▼				
	143	Krizova 2019	Genomic surveill	Krizova P; Honsk	10.1371/journal https://cloud.c		▼				
	146	Enkelmann 2020	Disentangling ou	Enkelmann J; voi	10.1017/S09502 https://cloud.c		▼				
	147	Javadi 2021	Real-Time Invest	Javadi W; Ehni J;	10.1093/cid/ciae https://cloud.c		▼				
	152	Dickinson 2019	Whole-genome	Dickinson H; Rea	10.1016/j.jhin.21 https://cloud.c		▼				
	153	Sherry 2019	Genomics for Mi	Sherry NL; Lane	10.1128/JCM.00 https://cloud.c		▼				
	160	Baier 2022	Target capture si	Baier C; Huang J	10.1186/s13756 https://cloud.c		▼				
	165	Houlihan 2018	Use of Whole-Ge	Houlihan CF; Fra	10.1093/infdis/fji https://cloud.c		▼				
	167	Kaku 2022	Changing molecu	Kaku N; Sasaki D	10.1093/jac/dka https://cloud.c		▼				
	174	Ugolotti 2018	Genomic charact	Ugolotti E; Di M	10.1016/j.jhin.21 https://cloud.c		▼				
	181	CremersAIH 202	Surveillance-eml	Cremers AIH; Co	10.1038/s41598 https://cloud.c		▼				
	185	Snell 2022	Combined epide	Snell LB; Fisher	10.1016/j.cml.21 https://cloud.c		▼				
	202	Fitzpatrick 2016	Utility of Whole	Fitzpatrick MA;	10.1128/JCM.01 https://cloud.c		▼				
	203	Kossow 2019	Whole genome	Kossow A; Kamp	10.1016/j.jhin.21 https://cloud.c		▼				

Note

CHECK: Make sure the "Covidence #" column and the "Data Storage Location" column are in the same order before merging.

Draft the "Metadata Legend" & "Metadata Table" tab

- 12 The **"Metadata Table"** will be for recording whether a contextual metadata field has been identified in a resource by the presence of data that could occupy it. We use the word "identify" because we are not collecting the data for use, but rather to indicate "TRUE" or "FALSE" for whether this data was collected by the researchers producing the resource. The goal is to eventually look at the correlation between the research performed and the contextual data fields they collected. Because this is a qualitative study, instead of simply putting "TRUE", the data value/quote that was used to justify "TRUE" is listed in its place. This way if there is eventually some debate about whether a study included a data field or not, we can understand where the data collector was coming from and even return to the original resource(s) to find the value/quote in context. **A "Metadata Type (Field Name)" is labelled as TRUE by the inclusion of the quoted "value" which supports the conclusion.**

Because some contextual metadata fields may be broader than others, "Subfields" may be worth collecting at a more granular level. In cases when such granular data isn't needed, "Example Subfields" may be included to help data extractors understand whether they have found a positive metadata value or not. **If collecting data on narrower subfields, the overarching "Metadata Type (Field Name)" are deemed "TRUE" if a required subfield was identified, or in the case where data matching an example subfield value was found.**

- 12.1 With this information in mind, we can begin drafting the "Metadata Legend" and "Metadata Table". You are going to want to **begin by deciding what kinds of metadata you are interested in exploring**. To help you define metadata fields consider the following one or more of the following resources:

- A contextual metadata expert or data specification developer
- Public Health Alliance for Genomic Epidemiology (PHA4GE) contextual data specifications (<https://github.com/pha4ge>)
- Centre for Infectious Disease and One Health (CIDGOH) contextual data specifications (<https://github.com/cidgoh>)
- Minimum information standards (e.g. MIxS: Minimum Information about any (X) Sequence; <https://github.com/GenomicsStandardsConsortium/mixs>)
- National Center for Biotechnology Information (NCBI) Biosample template and/or packages
- European Nucleotide Archive (ENA) data Standards
- DNA DataBank of Japan (DDBJ) data Standards
- GISAID metadata standards
- etc.

Another good place to look are the **Open Biological and Biomedical Ontologies (OBO) Foundry ontologies** which can be browsed on <https://www.ebi.ac.uk/ols4/> or <https://ontobee.org/>.

12.2 **Create the following headers in the "Metadata Legend" tab.** There will be two "parts" to this legend, which you can have side by side or separated out into different tabs (leaving them separate or recombining them in the final stages). One table explains the **"Metadata Table" "Headers/Fields"**, like we did for the "Overview Table", while the other expands on the "Metadata Type (Field Name)" specification information.

"Metadata Legend" Tab "Metadata Table - Headers":

- **Metadata Type (Field Name)** - Specific fields that are usually "required" when present in a specification.
- **Description / Definition** - Field or subfield definitions, pulled directly from specification vocabulary whenever possible with additional information added as needed.
- **Values** - Quote from the text the information used as evidence to conclude the presence of a metadata type.

Optional:

- **Motivations** - A positive or negative in-text motivation for the inclusion of a metadata type. A positive motivation could be along the lines of "<this metadata> allowed us to do this", while a negative motivation could be "a limitation of that study was we couldn't study this because we didn't have <this metadata>".

EXAMPLE

"Metadata Legend" Tab "Headers" Table:

A	B	C
Header	Description / Definition	version 0.1
Metadata Type (Field Name)	Metadata fields that are usually required when present in a CIDGOH specification.	
Definition	Field or subfield definitions, some pulled directly from specification vocabulary while others are more broadly described categories.	

	A	B	C
	Values	Quote from the text the information used as evidence to conclude the presence of a metadata type.	
	Motivation	A positive or negative motivation for the inclusion of a metadata type. A positive motivation would be along the lines of "this metadata allowed us to do this", while a negative motivation could be "a limitation of that study was we couldn't study this because we didn't have that".	

A simple version of the "Metadata Legend" "Headers" table where there is no hierarchy to the subfields.

"Metadata Table" Tab:

	A	B	C	D
	Metadata Type (Field Name)	Definition	#	#
	version 0.1		Values	Motivation

The initial layout of the "Metadata Table" tab. Do not include the "Motivation" column if not applicable. The "#"'s are place holders for the assigned "Covidence #". The "Values" and "Motivations" appear in pairs for the same "Covidence #". The version # can be placed somewhere visible but out of the way.

Note

Another column that can be added to support the "Values" column is a "Location" column that indicates what resource the "Value" was pulled from (e.g. manuscript, supplement, accession lookup) to make it easier to track down the value as needed. It could also be used to do a more comprehensive analysis on the availability (e.g. in-text, in-figure, supplementary) of individual data fields, but it adds to the labour of data extractors so it is only recommended if you are working with a small number of fields or a small number of resources undergoing extraction.

12.3 "Metadata Legend" Tab "Metadata Type (Field Name)" Table:

- **Metadata Type (Field Name)** - Specific fields that are usually "required" when present in a CIDGOH specification.

- **Description / Definition** - Field or subfield definitions, pulled directly from specification vocabulary whenever possible.
- **Example** - One or more examples of contextual metadata that could be extracted to satisfy the field.

EXAMPLE

"Metadata Legend" Tab "Metadata Type (Field Name)" Table:

	A	B	C	D
	Metadata Type (Field Name)	Description / Definition	Example	version 0.1

The initial layout of the "Metadata Type (Field Name)" table in the "Metadata Legend" tab. The addition of the "Description / Definition" and "Example" columns in this sheet are optional - they were added for this protocol because user testing found it helpful for many data extractors. Individual data collectors can easily "Freeze" the columns they would like to always maintain in view.

	A	B	C	D	E	F	G
1	Metadata Type (Field Name)	Description / Definition	Example	Header	Description / Definition	version 0.1	
2				Metadata Type (Field Name)	Metadata fields that are usu		
3				Definition	Field or subfield definitions,		
4				Values	Quote from the text the info		
5				Motivation	A positive or negative motiv		
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							

Empty "Metadata Legend" tab showing both the "Metadata Type (Field Name)" [LEFT] and "Metadata Legend - Headers" [RIGHT] in the same tab. This layout is useful to data extractors so they have a single point of reference for information on the "Metadata Table" tab, but it can be inconvenient while you are adding to the upper rows of the "Metadata Legend - Metadata Type (Field Name)".

	A	B	C	D	E	F	G
1	Metadata Type (Field Name)	Description / Definition	Example	Header	Description / Definition	version 0.1	
2	isolate ID	Identifier of the specific isolate.	SARS-CoV-2/human/USA/CA-CDPH-	Metadata Type (Field Name)	Metadata fields that are usu		
3	isolation date / isolate collection date	The date on which the isolate w	2023-08; 2023-08-23; April 5th, 202	Definition	Field or subfield definitions,		
4	isolate received date	The date on which the isolate w	2023-08; 2023-08-23; April 5th, 202	Values	Quote from the text the info		
5	isolated by	The name of the agency, organ	Public Health Agency of Canada (PH	Motivation	A positive or negative motiv		
6	geo_loc name (country)	The country where the sample v	South Africa, Australia, Bahamas, C				
7	geo_loc name (state/province/territory)	The state/province/territory wher	Western Cape; British Columbia; Prc				
8	geo_loc name (county/region)	The county/region of origin of th	Derbyshire, Orange County, Queens				
9	geo_loc name (city)	The city of origin of the sample.	Vancouver; Mumbai; Seoul; Dubai				
10	geo_loc name (site)	The name of a specific geographi	Credit River; Lake Ontario; Black Cr				
11	geo_loc latitude	The latitude coordinates of the g	38.98 N; 66° 34' (66.57°) N				
12	geo_loc longitude	The longitude coordinates of the	77.11 W; 0 degrees longitude (prime				
13	food product origin geo_loc name (country)	The country of origin of a food p	South Africa, Australia, Bahamas, C				
14	sample collected by	The name of the agency that col	Public Health Agency of Canada (PH				
15	sample collection date	The date on which the sample w	2023-08; 2023-08-23; April 5th, 202				
16	sample collection date precision	The precision to which the "sam	day (e.g. 2023-08-23); month (e.g. 2				
17							

"Metadata Legend" tab with "Metadata Type (Field Name)" data without "levels".

13 Managing "levels" of metadata fields. (OPTIONAL)

Some specifications may have levels of contextual metadata fields or you may conceive of groupings that you find of particular interest.

E.g. The CIDGOH specification fields "geo_loc name (city)" and "geo_loc latitude" are more granular than "geographical location (sample)" because they are both more specific versions of information that the latter encompasses.

For the purposes of this protocol, levels have been organized as follows:

- **Category/Bin** - A broad metadata category used to organize the collection template for ease of use of data extractors. They may be used as broad categories of analysis, but be sure to check other categories/bins for subfields that may also count towards this category (i.e., check for polyhierarchical relationships).
- **Field** - An "upper level" field, for which the presence of a "subfield" counts as a "TRUE" positive.
- **Subfield (Required)** - A "narrow"/"lower level" field that can count towards an "upper level" field and is deemed required for granular data collection.
- **Subfield (Example)** - A "narrow"/"lower level" field that can count towards an "upper level" field but is only present as an example to help data extractors identify a "TRUE" positive for the "upper level" (i.e., data is not being collected at a granular level for this field).

You could keep these broader field categories off of the data collection sheet and just organize and apply them to the groupings during data analysis. However, if you want to indicate the presence of a broader data field when an appropriate data value/quote is found that does not fit the criteria of your listed subfields, but does for the example subfields, you will want to proceed with this step.

- 13.1 If you are adding levels add the following to your "Metadata Legend" "Headers" Table and **update the "Description/Definitions" to appropriately reflect the levels.**

EXAMPLE

"Metadata Table" Tab:

A	B	C
Header	Description / Definition	version 0.1
Metadata Type (Field Name)	The broader metadata / contextual data field type / category.	
Required Subfields	Specific subfields that are usually "required" when present in a CIDGOH specification.	
Example Subfields	Subfields that would be organized under a specific type / category, but don't tend to be required (e.g. instead are recommended or additional options).	
Definition	Field or subfield definitions, some pulled directly from specification vocabulary while others are more broadly described categories.	
Values	Quote from the text the information used as evidence to conclude the presence of a metadata type.	
Motivation	A positive or negative motivation for the inclusion of a metadata type. A positive motivation would be along the lines of "this metadata allowed us to do this", while a negative motivation could be "a limitation of that study was we couldn't study this because we didn't have that".	

"Metadata Legend" "Headers" table with additional field levels and examples of new and updated "Description / Definition" information (shown in bold).

"Metadata Legend" Tab "Headers" Table:

A	B	C	D	E	F
Metadata Type (Field Name)	Required Subfield	Definition	Example Subfield	#	#
version 0.1				Values	Motivation

Adding required and example subfield columns (shown in bold). How you organize the addition of these columns is up to you. We recommend keeping the two field columns on the left so that data collectors can freeze them and then leave out the reference columns they don't always want in view.

EXAMPLE

	A	B	C	D	E	F	G
1	Metadata Type (Field Name)	Description / Definition	Example		Header	Description / Definition	version 0.1
2	isolate ID	Identifier of the specific isolate.	SARS-CoV-2/human/USA/CA-CDPH-		Metadata Type (Field Name)	The broader metadata / con	
3	isolation date / isolate collection date	The date on which the isolate w	2023-08; 2023-08-23; April 5th, 202		Required Subfields	Specific subfields that are u	
4	isolate received date	The date on which the isolate w	2023-08; 2023-08-23; April 5th, 202		Example Subfields	Subfields that would be orgi	
5	isolated by	The name of the agency, organi	Public Health Agency of Canada (PH		Definition	Field or subfield definitions,	
6	organism	Taxonomic name of the organisr	Severe acute respiratory syndrome (		Values	Quote from the text the infor	
7	purpose of sampling	The reason that the sample was	Diagnostic testing; Survey study; Pro		Motivation	A positive or negative motiv	
8	geo_loc name (country)	The country where the sample v	South Africa, Australia, Bahamas, C				
9	geo_loc name (state/province/territory)	The state/province/territory wher	Western Cape, British Columbia, Pro				
10	geo_loc name (county/region)	The county/region of origin of th	Derbyshire, Orange County, Queens				
11	geo_loc name (city)	The city of origin of the sample.	Vancouver, Mumbai, Seoul, Dubai				
12	geo_loc name (site)	The name of a specific geographi	Credit River, Lake Ontario, Black Cr				
13	geo_loc latitude	The latitude coordinates of the g	38.98 N; 66° 34' (66.57°) N				
14	geo_loc longitude	The longitude coordinates of the	77.11 W; 0 degrees longitude (prime				
15	food product origin geo_loc name (cour	The country of origin of a food p	South Africa, Australia, Bahamas, C				
16	sample collected by	The name of the agency that col	Public Health Agency of Canada (PH				
17	sample collection date	The date on which the sample w	2023-08; 2023-08-23; April 5th, 202				
18	sample collection date precision	The precision to which the "sam	day (e.g. 2023-08-23); month (e.g. 2				
19							

"Metadata Legend" tab with "Metadata Type (Field Name)" data without "levels".

	A	B	C	D	E	F	G
1	Metadata Type (Field Name)	Description / Definition	Example		Header	Description / Definition	version 0.1
2	Sample Collection				Metadata Type (Field Name)	The broader metadata / conte	
3	isolate metadata				Required Subfields	Specific subfields that are usu	
4	isolate ID	Identifier of the specific isolate	SARS-CoV-2/human/USA/C		Example Subfields	Subfields that would be organi	
5	isolation date / isolate collection date	The date on which the isolate	2023-08; 2023-08-23, April		Definition	Field or subfield definitions, sc	
6	isolate received date	The date on which the isolate	2023-08; 2023-08-23, April		Values	Quote from the text the inform	
7	isolated by	The name of the agency, orga	Public Health Agency of Ca		Motivation	A positive or negative motivati	
8	organism	Taxonomic name of the organi	Severe acute respiratory sy				
9	purpose of sampling	The reason that the sample w	Diagnostic testing; Survey s				
10	geographical location (sample)	Data which describes the loca					
11	geo_loc name (country)	The country where the sample	South Africa, Australia, Bah				
12	geo_loc name (state/province/territory)	The state/province/territory wh	Western Cape, British Colou				
13	geo_loc name (county/region)	The county/region of origin of	Derbyshire, Orange County				
14	geo_loc name (city)	The city of origin of the sampl	Vancouver, Mumbai, Seoul;				
15	geo_loc name (site)	The name of a specific geograp	Credit River; Lake Ontario; l				
16	geo_loc latitude	The latitude coordinates of the	38.98 N; 66° 34' (66.57°) N				
17	geo_loc longitude	The longitude coordinates of t	77.11 W; 0 degrees longituc				
18	food product origin geo_loc name (cour	The country of origin of a food	South Africa, Australia, Bah				
19							

"Metadata Legend" tab with "Metadata Type (Field Name)" data with the early addition of "levels" (yellow highlight).



1	type (temporary)	Metadata Type (Field Name)	Description / Definition	Example	Header	Description / Definition	version 0.1
2	Category/Bin	Sample Collection			Metadata Type (Field Name)	The broader metadata / conte	
3	Field	isolate metadata			Required Subfields	Specific subfields that are usu	
4	Subfield	isolate ID	Identifier of the specific isolate SARS-CoV-2/human/USA/C		Example Subfields	Subfields that would be organi	
5	Subfield	isolation date / isolate collection date	The date on which the isolate 2023-08; 2023-08-23; April		Definition	Field or subfield definitions, sc	
6	Subfield	isolate received date	The date on which the isolate 2023-08; 2023-08-23; April		Values	Quote from the text the inform	
7	Subfield	isolated by	The name of the agency, orga Public Health Agency of Ca		Motivation	A positive or negative motivati	
8	Field	organism	Taxonomic name of the organi Severe acute respiratory sy				
9	Field	purpose of sampling	The reason that the sample w Diagnostic testing; Survey s				
10	Field	geographical location (sample)	Data which describes the loca				
11	Subfield	geo_loc name (country)	The country where the sample South Africa; Australia; Bah				
12	Subfield	geo_loc name (state/province/territory)	The state/province/territory w Western Cape; British Colu				
13	Subfield	geo_loc name (county/region)	The county/region of origin of Derbyshire; Orange County				
14	Subfield	geo_loc name (city)	The city of origin of the sampl Vancouver; Mumbai; Seoul;				
15	Subfield	geo_loc name (site)	The name of a specific geogra Credit River; Lake Ontario; I				
16	Subfield	geo_loc latitude	The latitude coordinates of the 38 98 N; 66° 34' (66.57°) N				
17	Subfield	geo_loc longitude	The longitude coordinates of t 77 11 W; 0 degrees longituc				
18	Subfield	food product origin geo loc name (cour	The country of origin of a food South Africa; Australia; Bah				
19							

"Metadata Legend" tab with a column to type the different "Metadata Type (Field Name)" data "levels" (yellow highlight). This is a useful format for downstream data manipulation but is not the best user experience for data extractors.

1	A	B	C	D	E	F	G	H
1	Metadata Type (Field Name)	Subfield	Description / Definition	Example	Header	Description / Definition	version 0.1	
2	Sample Collection				Metadata Type (Field Name)	The broader metadata / contextual data fi		
3	isolate metadata				Required Subfields	Specific subfields that are usually "require		
4		isolate ID	Identifier of the specific isolate SARS-CoV-2/human/USA/C		Example Subfields	Subfields that would be organized under i		
5		isolation date / isolate collection date	The date on which the isolate 2023-08; 2023-08-23; April		Definition	Field or subfield definitions, some pulled c		
6		isolate received date	The date on which the isolate 2023-08; 2023-08-23; April		Values	Quote from the text the information used i		
7		isolated by	The name of the agency, orga Public Health Agency of Ca		Motivation	A positive or negative motivation for the ir		
8	organism		Taxonomic name of the organi Severe acute respiratory sy					
9	purpose of sampling		The reason that the sample w Diagnostic testing; Survey s					
10	geographical location (sample)		Data which describes the loca					
11		geo_loc name (country)	The country where the sample South Africa; Australia; Bah					
12		geo_loc name (state/province/territory)	The state/province/territory w Western Cape; British Colu					
13		geo_loc name (county/region)	The county/region of origin of Derbyshire; Orange County					
14		geo_loc name (city)	The city of origin of the sampl Vancouver; Mumbai; Seoul;					
15		geo_loc name (site)	The name of a specific geogra Credit River; Lake Ontario; I					
16		geo_loc latitude	The latitude coordinates of the 38 98 N; 66° 34' (66.57°) N					
17		geo_loc longitude	The longitude coordinates of t 77 11 W; 0 degrees longituc					
18		food product origin geo loc name (cour	The country of origin of a food South Africa; Australia; Bah					
19								

"Metadata Legend" tab with the temporary type column removed, and instead the field levels broken up across the upper level "Metadata Type (Field Name)" column (yellow highlight) and the lower level "Subfield". If you'd like, you can keep the "type" column in this template and not include it when importing the template to individual data collector sheets.

1	A	B	C	D	E	F	G	H
1	Metadata Type (Field Name)	Subfield (Required in Bold)	Description / Definition	Example	Header	Description / Definition	version 0.1	
2	Sample Collection				Metadata Type (Field Name)	The broader metadata / contextual data fi		
3	isolate metadata				Required Subfields	Specific subfields that are usually "require		
4		isolate ID	Identifier of the specific isolate SARS-CoV-2/human/USA/C		Example Subfields	Subfields that would be organized under i		
5		isolation date / isolate collection date	The date on which the isolate 2023-08; 2023-08-23; April		Definition	Field or subfield definitions, some pulled c		
6		isolate received date	The date on which the isolate 2023-08; 2023-08-23; April		Values	Quote from the text the information used i		
7		isolated by	The name of the agency, orga Public Health Agency of Ca		Motivation	A positive or negative motivation for the ir		
8	organism		Taxonomic name of the organi Severe acute respiratory sy					
9	purpose of sampling		The reason that the sample w Diagnostic testing; Survey s					
10	geographical location (sample)		Data which describes the loca					
11		geo_loc name (country)	The country where the sample South Africa; Australia; Bah					
12		geo_loc name (state/province/territo	The state/province/territory w Western Cape; British Colu					
13		geo_loc name (county/region)	The county/region of origin of Derbyshire; Orange County					
14		geo_loc name (city)	The city of origin of the sampl Vancouver; Mumbai; Seoul;					
15		geo_loc name (site)	The name of a specific geogra Credit River; Lake Ontario; I					
16		geo_loc latitude	The latitude coordinates of the 38 98 N; 66° 34' (66.57°) N					
17		geo_loc longitude	The longitude coordinates of t 77 11 W; 0 degrees longituc					
18		food product origin geo loc name (cour	The country of origin of a food South Africa; Australia; Bah					
19								

"Metadata Legend" tab with the field levels broken up across the upper level "Metadata Type (Field Name)" column (yellow highlight) and the lower level "Subfield". Required fields are indicated in bold.

13.2 "Metadata Table" formatting with "levels".

EXAMPLE

Metadata Type (Field Name):

- **Category/Bin** - No additional variable data.
- **Field** - Has example subfields and ideally a description.

Subfield:

- **Subfield (Required)** - bold
- **Subfield (Example)** - not in bold

Metadata Type (Field Name) version 8	Required Subfield	Definition	Example Subfield
Sample Collection			
isolate metadata	<i>Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE</i>	Data which describes an isolate.	isolation date / isolate collection date isolate received date isolated by isolate database accession ID
	Isolate ID	Identifier of the specific isolate.	
organism	<i>Quote value if example found.</i>	Taxonomic name of the organism.	
purpose of sampling	<i>Quote value if example found.</i>	The reason that the sample was collected.	
geographical location (sample)	<i>Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE</i>	A data field which describes the location of origin of the sample.	geo_loc name (county/region) geo_loc name (city) geo_loc name (site) geo_loc latitude geo_loc longitude food product origin geo loc name (country)
	geo_loc name (country)	The country where the sample was collected.	
	geo_loc name (state/province/territory)	The state/province/territory where the sample was collected.	
sample collected by	<i>Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE</i>	Data which describes the information regarding who performed the sample collection and how.	
	sample collected by	The name of the agency that collected the original sample.	
	sample collection date	The date on which the sample was collected.	
	sample collection date precision	The precision to which the "sample collection date" was provided. (i.e. the precision of granularity to the "day", "month", or "year" for the date provided in the "sample collection date" field.)	

"Metadata Table" tab that reflects the different levels created in the "Metadata Legend" tab.

14 **CHECK: Do you need to break a field down into subfields? Are concepts being merged together?**

Note

EXAMPLE: Is "Environmental Site" described in a way that allows for data entry that isn't a location? Or is the example provided by the original specification for an "Environmental Material"? Or perhaps the "Environment" field should be broken down into "Environmental Site" and "Environmental Material" levels for more granular analysis?

15 **Review with your data extraction team and advisors.**

16 **Create the "Metadata Table" collection template in the "Metadata Table" tab.**

- The headers should be the same as those listed in the "Metadata Legend"; depending on how many you have you can choose to switch the table orientation. If you have a lot of contextual metadata fields (the examples in this protocol do), it is best to keep a "horizontal" table format for data collectors. However, had there been fewer fields (e.g. <15) we would have gone with a "vertical" table layout.
- Don't forget your version information.
- "Freeze" the field columns as default so that they remain in view as users enter data. Users may end up wanting to change the freeze to include the descriptive columns as well - this will not cause an issue with permissions or imports.
- If you want you can add additional formatting.

Note

If you want to save yourself from having to update both the "Metadata Legend" and "Metadata Table" separately you can use the Sheets "=IMPORTRANGE()" function to copy the contents over from "Metadata Legend" into the "Metadata Table" tab, adding "=TRANPOSE" if you want a different orientation. This may be the best way to go for a smaller and/or non-levelled set of metadata fields.

We did not do this because user testing decided on a different layout (see EXAMPLE below) for "Example Subfield" and reminder guidance was added (light grey italic text) to help data collectors implement the protocol. These changes had to be applied directly to cell values which would be an issue if using the "=IMPORTRANGE()" function because editing cell values causes the imported values to fail/vanish. More on this under the "Dynamic Spreadsheets" section.

EXAMPLE

	A	B	C	D	E	F
1	Metadata Type (Field Name)	Required Subfield	Definition	Example Subfield	#	#
2	version 8				Values	Motivation
3	Sample Collection					
4	isolate metadata	Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	Data which describes an isolate.	isolation date / isolate collection date isolate received date isolate database accession ID		
5		Isolate ID	Identifier of the specific isolate.			
6	organism	Quote value if example found	Taxonomic name of the organism.			
7	purpose of sampling	Quote value if example found	The reason that the sample was collected.			
8	geographical location (sample)	Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	A data field which describes the location of origin of the sample.	geo_loc name (county/region) geo_loc name (city) geo_loc name (site) geo_loc latitude geo_loc longitude food product origin geo_loc name (country)		
9		geo_loc name (country)	The country where the sample was collected.			
10		geo_loc name (state/province/territory)	The state/province/territory where the sample was collected.			
11	sample collected by	Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	Data which describes the information regarding who performed the sample collection and how.			
12		sample collected by	The name of the agency that collected the original sample.			
13		sample collection date	The date on which the sample was collected.			
14		sample collection date precision	The precision to which the "sample collection date" was provided. (i.e. the precision of granularity to the "day", "month", or "year" for the date provided in the "sample collection date" field.)			
15	sample identifiers	Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	Identifiers used to label the sample. A sample/isolate may have multiple identifiers.			
16		specimen collector sample ID	The user-defined name for the sample.			
17		sample database accession ID	Alphanumeric code assigned in repository of data for the sample.			

Example "Metadata Table" tab ready for assignment with user guidance added in light grey (these are optional and were added post-pilot data extraction based on user-experience feedback).

Note

FORMATTING RECOMMENDATION: Adding conditional formatting in data input areas can help the users identify "at a glance" values they missed checking and need to do a final review on to confirm presence or absence. We went with a dark "tone" (which works with colour blindness) when "Cell is empty" and "Text is exactly 'FALSE'" (optionally). If you apply this formatting at this stage you can then copy & paste to apply it to your templates.

Preparing files for contextual metadata extraction

- 17 We found a helpful step in preparation for contextual metadata extraction was to **get copies of all the documents/resources we were intending to review downloaded and organized within a cloud service before extraction**. This was especially useful for closed-access manuscripts on the rare occasion where one data extractor had credentials for access while another did not or was at least unsuccessful. You may need lab IT support to execute this step.



We recommend Nextcloud and the guidance that follows will be specific to working with that platform.

Software

Nextcloud

NAME

<https://nextcloud.com/>

SOURCE LINK

From our experience, the COVIDENCE API supports the bulk upload, but not bulk download, of manuscript PDFs. So a custom python script was written to download the manuscripts from the available DOIs and then upload them to our Nextcloud server (available below). If this is a barrier for you, you can continue to the next substep (16.1) for ideas on proceeding manually, or you can explore whether your existing (or other) reference manager has the capacity to download the majority of the PDFs for you.



scoping_review_pdf_download_and...

```
import pandas as pd
import subprocess
import numpy as np
import os

# Read the data from the tab-separated file (adjust the file path
as needed)
from urllib.parse import urljoin

base_url =
'https://cloud.name.ca/remote.php/dav/files/dj/project/test/'

credential = "username:password"

data = pd.read_csv("input2.txt", sep='\t')

# Iterate through the rows of the DataFrame
for index, row in data.iterrows():
    doi = row["DOI"]
    covidence_number = row["Covidence #"]
    covidence_number_only = covidence_number.replace("#", "")

    # Check if DOI is missing (NaN)
    if pd.isna(covidence_number):
        # Construct the scidownl command
        command = f"scidownl download --doi {doi} --out
./pdf_results/{covidence_number_only}.pdf"
        url = urljoin(base_url, covidence_number_only)
        command1 = f"curl -X MKCOL {url} -u {credential}"

        # Execute the command
        try:
            subprocess.run(command1, shell=True, check=True)
            print(f"Create folder for: {doi} and Covidence #:
{covidence_number} completed successfully.")
        except subprocess.CalledProcessError as e:
            print(f"Error for DOI: {doi} and Covidence #:
{covidence_number}: {e}")

    # Upload file
    file_path = "pdf_results/"+covidence_number_only+".pdf" #
Replace with the actual file path
    print(file_path)
    # Check if the file exists
```

```
if os.path.exists(file_path):
    url1 = urljoin(base_url, covidence_number_only+"/")
    print(url1)
    url2 = urljoin(url1, covidence_number_only+".pdf")
    print(f"The file '{file_path}' exists. Now trying to
upload file")
    command2 = f"curl -T {file_path} {url2} -u
{credential}"

    try:
        subprocess.run(command2, shell=True, check=True)
        print(f"Uploading fulltext in pdf format: {doi}
and Covidence #: {covidence_number} completed successfully.")
    except subprocess.CalledProcessError as e:
        print(f"Uploading Error for DOI: {doi} and
Covidence #: {covidence_number}: {e}")

else:
    print(f"The file '{file_path}' does not exist")

else:
    print(f"Skipping line {index + 1} because covidence number
is missing (NaN).")
```

- 17.1 Once the script has successfully ran, **data extractors need to review the contents of the folders for studies they were assigned; checking to resolve if files were missing, corrupted, or not the appropriate manuscript.** Some may also want to use this time to download additional supplemental data and sample accession data (one random sample if available), rather than waiting to do this during extraction.

1. Go to your assigned "Overview Table"
2. *Navigate* to the "Date Storage Location" column for the study
3. *Select* the URL

If a PDF is present:

4. *Open* the PDF and double check it matches the study information in the overview table

If there isn't a PDF present or the PDF is corrupted:

- **Option 1)** Using COVIDENCE:

All missing PDFs would've been uploaded to COVIDENCE during the abstract and full manuscript review.

4. Go into COVIDENCE
5. Search for the study COVIDENCE number (do not include "#")
6. Download the study PDF
7. Upload PDF to your data cloud
8. **OPTIONAL**: Proceed to "**Option 2) Steps 4-6**" for supplemental data

- **Option 2)** *not* using COVIDENCE / Obtaining supplementary data

4. Use the DOI and other study information from the "Overview Table" to locate the article
5. If you cannot access the article through your credentials, *request* assistance from a work peer or superior
6. **OPTIONAL**: Use this time to *download* and *upload* any supplemental and/or accession data you will be using for data extraction

Note

LIMITATION: It is resource exhaustive to extract metadata from more than one sample accession, or all of them, but by not doing so there could be cases where the sample selected has more or fewer fields than other samples. Due to resource exhaustion, we selected one sample accession at random to extract metadata from.

17.2 Generate Nextcloud URLs for the "Overview Table"

Fortunately, the way Nextcloud creates subfolder URLs makes it easy to generate "**Data Storage Location**" URLs since you only need to append "<folder name>" to the end of the link.

1. *Open* a new spreadsheet and *paste* in the "**Study/Covidence #**" column (EXAMPLE COLUMN B).
2. *Make* a new column for the "**Nextcloud Parent Folder URL**", it can be to the left or right of the previous column (EXAMPLE COLUMN A).
3. Use the "**=CONCATENATE()**" function, or equivalent, to add the contents of the two columns together (EXAMPLE COLUMN C); ensure that the order is "**Nextcloud Parent Folder URL**" followed by "**Study/Covidence #**".
4. The formula will *convert* to what you see in "EXAMPLE" "COLUMN D".
5. Double *check* that the order of the "**Study/Covidence #**" values are the same in both your current worksheet and the destination "**Overview Table**".

6. Copy the **"Results"** column contents.
7. Paste as values in the **"Data Storage Location"** column of your **"Overview Table"** (if you do a regular *paste*, the formula will not retain the values you made).

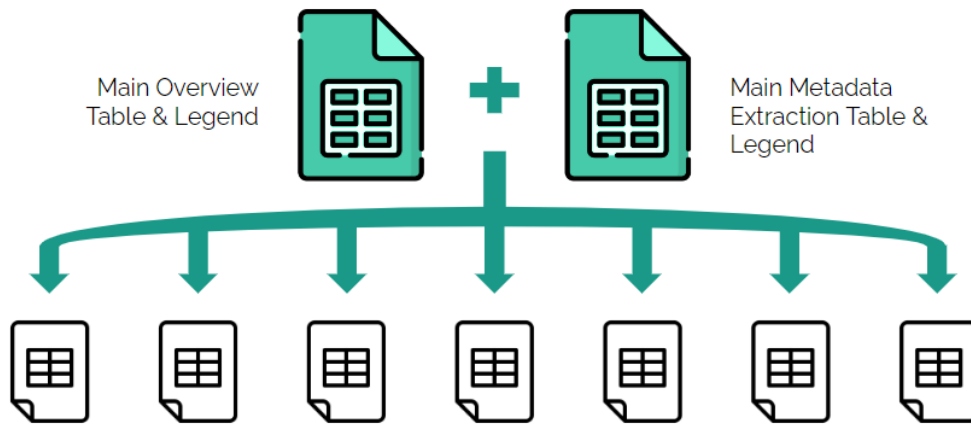
EXAMPLE

A	B	C	D
Nextcloud Parent Folder URL	Study/Covidence #	Formula (remove "'")	Result
https://cloud.ca/s/ABC?path=%2F	1	=CONCATENATE(A2,B2)	https://cloud.ca/s/ABC?path=%2F1
https://cloud.ca/s/ABC?path=%2F	2	=CONCATENATE(A3,B3)	https://cloud.ca/s/ABC?path=%2F2
https://cloud.ca/s/ABC?path=%2F	3	=CONCATENATE(A4,B4)	https://cloud.ca/s/ABC?path=%2F3

Generating "Data Storage Location" URLs using a spreadsheet application.

Dynamic Spreadsheets

- 18 **Begin building your collection templates.** Collection template tables are designed so that they import reference information from a main/master sheet; when the reference sheet is updated the changes appear on individual data collection sheets. This keeps data collectors in sync when it comes to the table legends, descriptions, examples, etc. These imported cells are protected so that only the project lead and/or the methods lead/manager can edit them to avoid users (i.e. data extractors) from accidentally breaking a formula.



Dynamic spreadsheets sending protected reference information from a "Main/Master/Reference" version to individual data extractor templates.

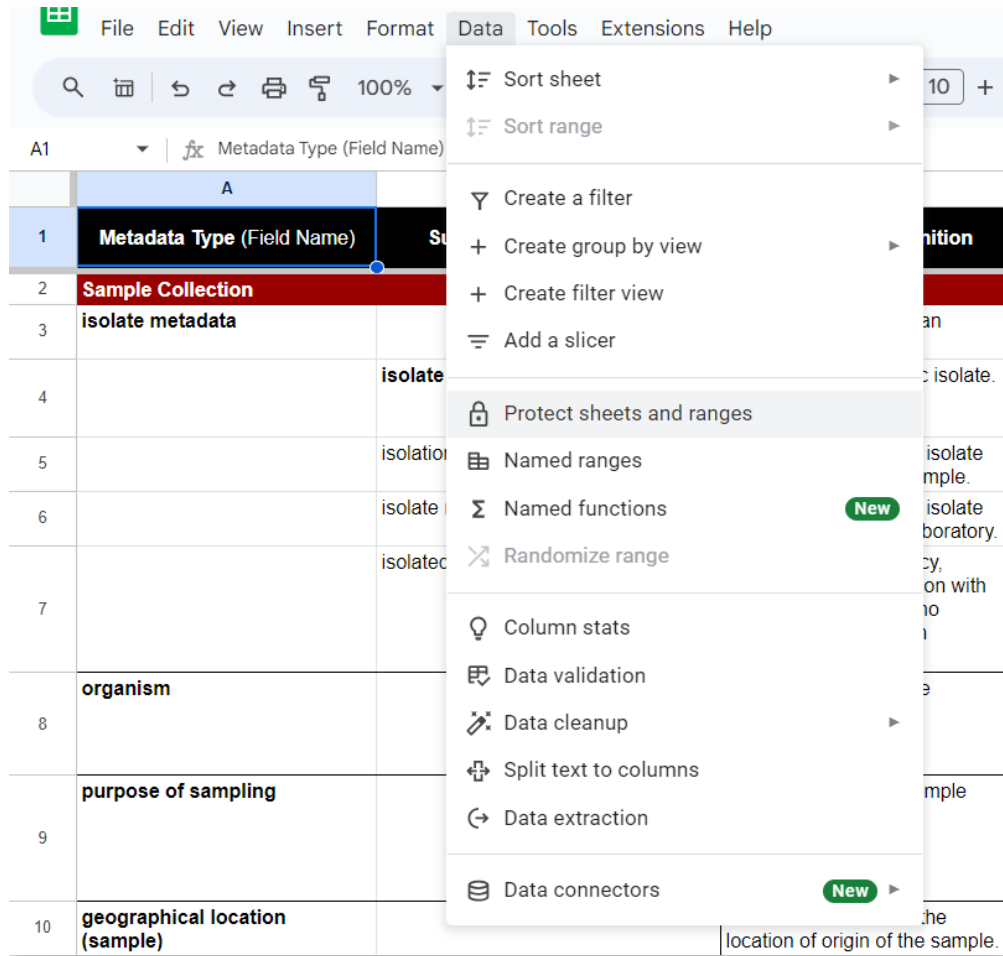
Note

BEFORE YOU BEGIN: If you are the only data collector you do not need to proceed with dynamic spreadsheets and can instead use your reference as the master.

Note

CONSIDER RECRUITING MORE DATA EXTRACTORS: One of the benefits of having a diverse group of data extractors is that it leads to engaging discussions on what qualifies for different types of data extraction and how to define the boundaries. Assumptions that may not have felt needed (e.g. obvious or overlooked) to be verbalized by experienced data extractors become drawn into the open and examined.

- 18.1 While in your **"Main/Master/Reference"** worksheet in Google Sheets go to "Data" in the upper ribbon and select **"Protect sheets and ranges"** from the dropdown.



- 18.2 Add a **"Description"** that clarifies what this protection is, *select* the **"Sheet"** option, and then **"Set permissions"**.



 Protected sheets & ranges ×

Range

Sheet

Metadata Legend (Master)

▼

☐ Except certain cells

Cancel

Set permissions

- 18.3 Select "**Restrict who can edit this range**" and then "**Only you**" or "**Custom**" if adding an additional team member (it is recommended to always have a backup access option).



Range editing permissions



☐ Show a warning when editing this range

☒ Restrict who can edit this range

Custom



Choose who can edit



Methods Manager



Project Lead



Add editors:

Enter names or email addresses

Done

18.4 **Bonus protection:** Repeat setting permissions for your sheet but instead of "**Restrict who can edit this range**" select "**Show a warning when editing this range**" so that your main editors get a reminder that changing this range impacts other sheets.

18.5 **Repeat the same steps for all tabs.** By the end you should have protected the following tabs:

- README
- Overview Legend
- Overview Table
- Metadata Legend
- Metadata Table

Note

TIP: When you get to "Range editing permissions" after the first time through, instead of selecting "Custom" under "Restrict who can edit this range", you can select "Copy permissions from another range".

- 19 **Now we are going to prepare a "blank template" worksheet with the imported reference material.** This template can then be duplicated for data collector assignment during the pilot and full data extraction. This is a completely new worksheet, not a new tab sheet in the "Main/Master/Reference" worksheet.

We will be using the Sheets "**=IMPORTRANGE()**" function as follows:

```
=IMPORTRANGE("google sheets URL", "'Tab Name'!Range-from:Range-to")
```

- 19.1 To get your "**google sheets URL**" go to the main reference sheet URL and select everything that appears before **"/edit?gid=..."**.

```
=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjqvzXA92DE-ofv27QFWZQ20pjdfg7rk", "'Tab Name'!Range-from:Range-to")
```



Google Docs URL with the relevant portion highlighted for use within the **=IMPORTRANGE()** function.

- 19.2 Now reference the name of tab in **"'Tab Name'"** that you wish to import from, e.g. the **"README"** tab.

```
=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjqvzXA92DE-ofv27QFWZQ20pjdfg7rk", "'README'!Range-from:Range-to")
```

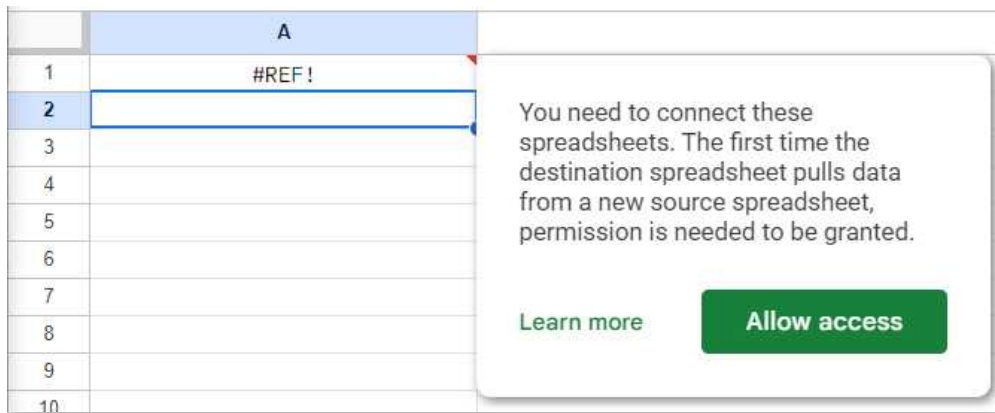
- 19.3 Specify the range in **"!Range-from:Range-to"** that you want imported, e.g. **"!A1:A"** for all of column A, **"!A1:B"** for all of column A and B, **"!A1:B14"** for column A and B down to row 14, etc.

```
=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjqvzXA92DE-0fV27QFWZQ20pjdfig7rk", "'README'!A1:B14")
```

Note

REMEMBER: Importing an entire column will slow down your worksheet. In many ways importing the entire column is more convenient to work with because when you add rows you don't have to go into the individual worksheets and add to the import range, but if you are noticing a slowdown or if you know it is unlikely you will be adding more rows, we recommend indicating a terminal row.

- 19.4 Since we are importing from another worksheet, not simply another tab within the same worksheet, you will need to **"Allow access"** to that sheet to ensure the data is pulled in.

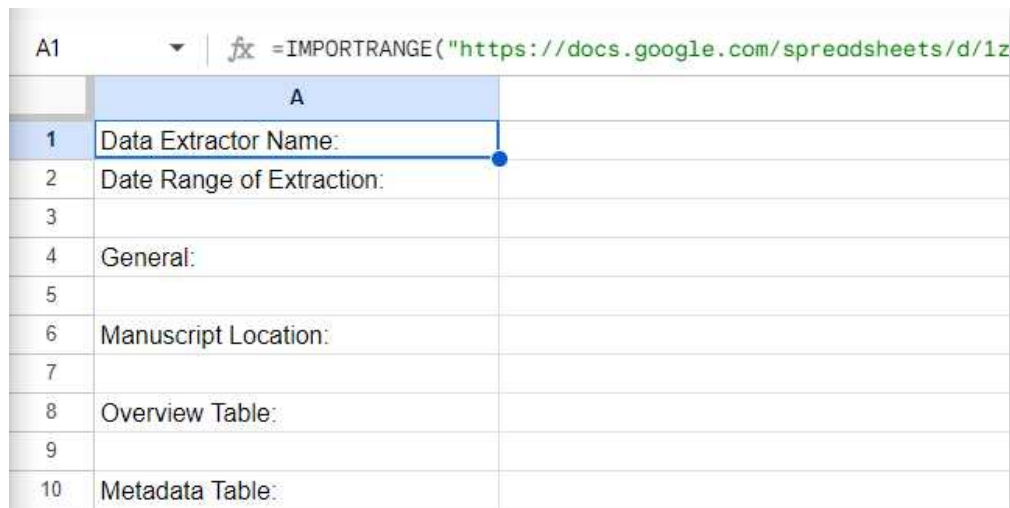


The "Allow access" warning that shows when you hover over the "#REF!" error that results from using =IMPORTRANGE() to pull from a different worksheet.

- 19.5 Add cell protections by navigating to **"Data"** > select **"Protect sheets and ranges"** > select **"Add a sheet or range"** > select **"Range"** > indicate which cells you would like to permit and then proceed with the **"Set permissions"** described above.

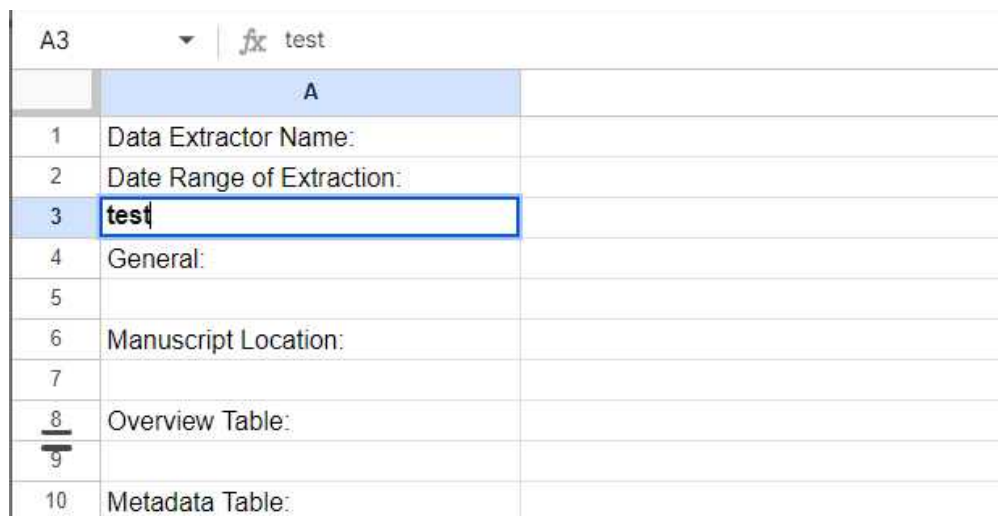
If you want to keep it simple you can just protect the =IMPORTRANGE() cell to prevent people from breaking the formula, but the import range will also get a "#REF!" error if any of the cells it is importing data over is edited. So, by protecting the entire range you will be less likely to have confused users who accidentally input a value in a cell.

EXAMPLE ERROR



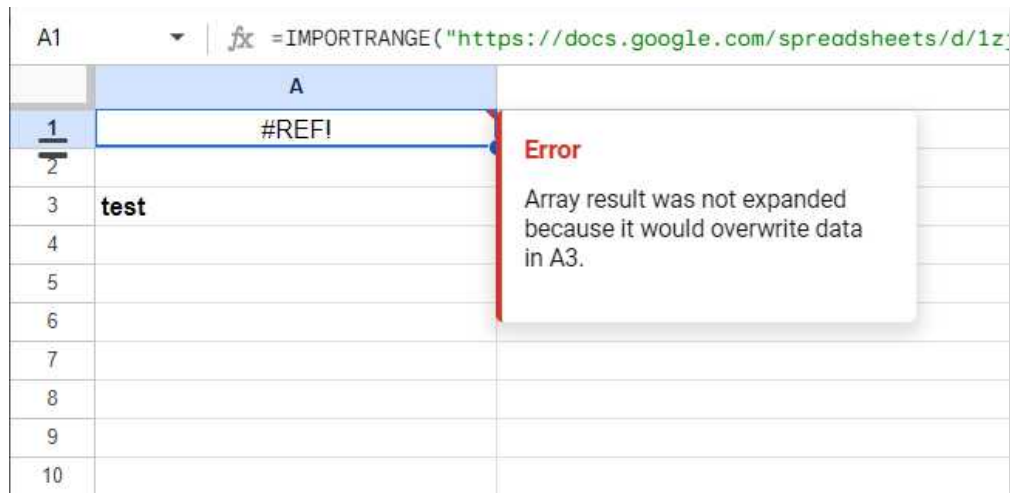
	A
1	Data Extractor Name:
2	Date Range of Extraction:
3	
4	General:
5	
6	Manuscript Location:
7	
8	Overview Table:
9	
10	Metadata Table:

Importing as expected.



	A
1	Data Extractor Name:
2	Date Range of Extraction:
3	test
4	General:
5	
6	Manuscript Location:
7	
8	Overview Table:
9	
10	Metadata Table:

Adding the text "test" to a cell within the range of the cells the =IMPORTFUNCTION() outputs over.



	A
1	#REF!
2	
3	test
4	
5	
6	
7	
8	
9	
10	

Getting a #REF! error because "test" text in A3 is overwriting the imported data.

20 Import "README" Template.

Note

In the future you may want to do the "Apply formatting" step before the "Import" step, but since it is optional and to avoid confusion it is being listed after this process.

20.1 =IMPORTRANGE() applied to column A. Import column A (applicable range).

EXAMPLE FUNCTION:

```
=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjqvzXA92DE-hbXiDsc0b_ofV27QFWZQ20pjdfg7rk", "'README'!A1:A14")
```



A1 fx =IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjqvzXA92DE-")

	A	B
1	Data Extractor Name:	
2	Date Range of Extraction:	
3		
4	General:	
5		
6	Manuscript Location:	
7		
8	Overview Table:	
9		
10	Metadata Table:	
11		
12	Metadata Table - Motivations:	
13		
14	Formatting/Customization:	

=IMPORTRANGE() applied to column A.

20.2 Import column B (applicable range); make sure you exclude the cells that you want your users to be able to input data into.

B4 fx =IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjqvzXA92DE-")

	A	B
1	Data Extractor Name:	
2	Date Range of Extraction:	
3		
4	General:	This is your personal worksheet for the scoping review full-te
5		
6	Manuscript Location:	You will find the link to the manuscript you've been assigned
7		
8	Overview Table:	The "Overview Table" is a summary table that provides data
9		
10	Metadata Table:	The "Metadata Table" is where more granular metadata evid
11		
12	Metadata Table - Motivations:	Another data element being collected are "quotes"/"evidence
13		
14	Formatting/Customization:	To make the data extraction worksheet more comfortable for

=IMPORTRANGE() applied to column B below the cells that users are expected to edit.

20.3 Cells highlighted to apply editing protections. Add placeholder text and editing protections.

	A	B
1	Data Extractor Name:	<First Name> <Last Name>
2	Date Range of Extraction:	<YYYY-MM-DD> to <YYYY-MM-DD>
3		
4	General:	This is your personal worksheet for the scoping review full-t
5		
6	Manuscript Location:	You will find the link to the manuscript you've been assigne
7		
8	Overview Table:	The "Overview Table" is a summary table that provides data
9		
10	Metadata Table:	The "Metadata Table" is where more granular metadata evi
11		
12	Metadata Table - Motivations:	Another data element being collected are "quotes"/"evidenc
13		
14	Formatting/Customization:	To make the data extraction worksheet more comfortable fo

Cells highlighted to apply editing protections. Column B1 & B2 (placeholder text cells) are not selected for editing protections because users need to be able to edit this information.

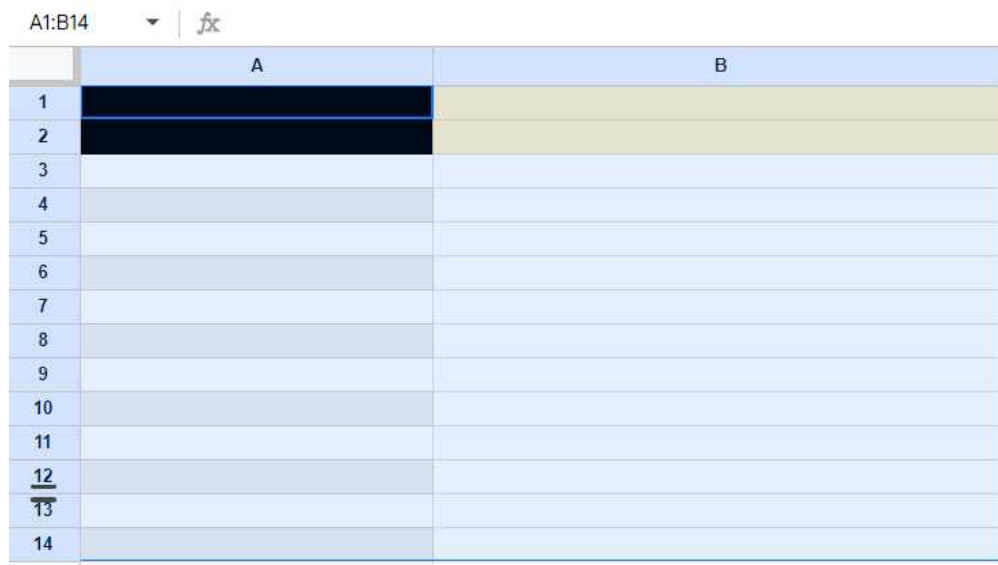
21 Apply formatting to the "README" tab.

21.1 README with custom formatted text/cells. Copy the formatted "Reference" version and paste it (with formatting) into a new tab.

A1	fx	Data Extractor Name:
	A	B
1	Data Extractor Name:	<First Name> <Last Name>
2	Date Range of Extraction:	<YYYY-MM-DD> to <YYYY-MM-DD>
3		
4	General:	This is your personal worksheet for the scoping review full
5		
6	Manuscript Location:	You will find the link to the manuscript you've been assigne
7		
8	Overview Table:	The "Overview Table" is a summary table that provides da
9		
10	Metadata Table:	The "Metadata Table" is where more granular metadata ev
11		
12	Metadata Table - Motivations:	Another data element being collected are "quotes"/"evidenc
13		
14	Formatting/Customization:	To make the data extraction worksheet more comfortable f

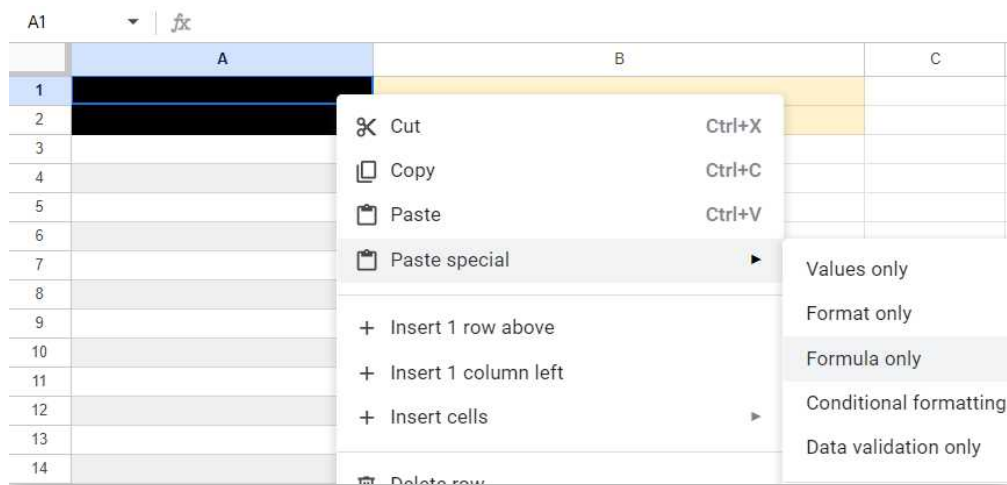
README with custom formatted cells.

21.2 Delete all cell values (do not clear the formatting).



README with custom formatted cells with the text contents deleted.

21.3 Copy your "Import column A" formula and paste special "Formula only".



"Formula only" paste option within the right click menu.



A1	A	B
1	Data Extractor Name:	
2	Date Range of Extraction:	
3		
4	General:	
5		
6	Manuscript Location:	
7		
8	Overview Table:	
9		
10	Metadata Table:	
11		
12	Metadata Table - Motivations:	
13		
14	Formatting/Customization:	

Text will display with the formatting preserved within the cells.

21.4 Copy your "Import column B" formula and *paste special* "Formula only".

B4	A	B
1	Data Extractor Name:	
2	Date Range of Extraction:	
3		
4	General:	This is your personal worksheet for the scoping review full-te
5		
6	Manuscript Location:	You will find the link to the manuscript you've been assigned
7		
8	Overview Table:	The "Overview Table" is a summary table that provides data
9		
10	Metadata Table:	The "Metadata Table" is where more granular metadata evic
11		
12	Metadata Table - Motivations:	Another data element being collected are "quotes"/"evidenc
13		
14	Formatting/Customization:	To make the data extraction worksheet more comfortable fo

"Formula only" pasted contents.

21.5 Add your placeholder values and apply text wrap formatting (if applicable).

E10		
	A	B
1	Data Extractor Name:	<First Name> <Last Name>
2	Date Range of Extraction:	<YYYY-MM-DD> to <YYYY-MM-DD>
3		
4	General:	This is your personal worksheet for the scoping review full-text metadata extraction. All headers are imported from a master reference sheet that will be updated as necessary to ensure everyone has the most up-to-date and comprehensive information available. It is recommended you familiarize yourself with and frequently reference the "Legend" tabs, as they are your guide to filling in the tables. All data extraction tables have a aforementioned "Legend" that describes headers, and provides value examples/descriptions of what those fields are for. Tables are designed so that they import reference information from a reference sheet; when the reference sheet is updated the changes appear on individual data collection sheets. This keeps data collectors in sync when it comes to the table "Legends", headers, etc. These imported cells are protected so that only the project lead (<Full Name>) or the methods manager (<Full Name>) can edit them to avoid users from accidentally breaking a formula. For data extraction examples, we recommend you review the information provided on your personal scoping review full-text metadata extraction pilot worksheet. All data extraction is to be performed manually as described in the protocol.
5		
6	Manuscript Location:	You will find the link to the manuscript you've been assigned under the "Nextcloud (Data Storage Location)" column. Make sure you double check that the title/date in the file matches the information in the overview table. If the file is missing, corrupted, or doesn't appear to be a manuscript, login to Covidence and search for your article using the "Covidence #" with the "#" sign in front.

README tab with placeholder text and formatting.

21.6 Delete the non-formatted tab you took your formulas from.

Note

MOVING FORWARD you do not need to create a second tab to move your formula over with, this was just to help you avoid having any errors on your first attempt.

22 Import "OVERVIEW LEGEND" Template.

22.1 Apply formatting (Optional): Copy the formatted "Reference" version and paste it (with formatting) into a new tab and then delete all cell values (do not clear the formatting).



G17				
	A	B	C	D
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				

"Overview Legend" with formatted cells, cell contents deleted.

22.2 Import columns (applicable range, don't forget your version information).

EXAMPLE:

```
=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjzqvzXA92DE-hbXiDscOb_ofV27QFWZQ20pjdfg7rk", "'Overview Legend'!A1:E38")
```

A1					
	A	B	C	D	E
1	Header	Description	Values	Value Description	Version 1
2	Assignee	The team member assigned to extra metadata from the associated article.			
3	Covidence #	The Covidence document reference number (#).			
4	Study	Shorthand study label indicating <1st author> <year>.			
5	Title	Manuscript/Document title.			
6	Authors	Manuscript/Document author list.			
7	DOI	DOI reference ID or PURL.			
8	Nextcloud (Data Storage Location)	URL link to the data storage folder on the Centre for Infectious Disease Genomics and One Health (Lab) Nextcloud server. This is where manuscripts can be found/uploaded, and where to upload additional data such as supplementary files and sequence data (when applicable).			
9	Variables of Interest	Overarching variables we've chosen to extract for analysis.			
10	Sequence Data Accessible (Database Name)	If the sequence data is publically available and accessible through information provided by the article, indicate which database it is located at.	<Database Name>	The name of the database where the sequence data was located.	
11			FALSE	Database information could not be identified.	
12	Sequence Data Saved (TRUE/FALSE)	Please save sequence information on the Lab server space dedicated for this purpose. Store in a folder labelled with the manuscript/document Covidence number.	TRUE	Sequence data or information about it has been stored on the Lab server.	
13			FALSE	Sequence data or information about it has not been stored on the Lab server.	

"Overview Legend" with formatted cells, cell contents imported.

22.3 Protect non-editable import cells / ranges.

[illegible]

"Overview Legend" with formatted cells, cell contents imported and protected from editing.

23 Import "METADATA LEGEND" Template:

23.1 **Apply formatting** (Optional): Copy the formatted "Reference" version and paste it (with formatting) into a new tab and then delete all cell values (do not clear the formatting).

G37	A	B	C	D	E	F
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						

"Metadata Legend" with formatted cells, cell contents deleted.

23.2 Import columns (applicable range, don't forget your version info).

EXAMPLE:

```
=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjqvzXA92DE  
-hbXiDsc0b_ofV27QFWZQ20pjdfg7rk", "'Metadata Legend'!A1:H108")
```

A1	=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjqvzXA9ZDE-hbXiDscOb_ofV27QFWZG20pjdFgTrk", "Metadata Legend!A1:H108")				
	A	B	C	D	E
1	Metadata Type (Field Name)	Subfield (Required in Bold)	Description / Definition	Example	Header
2	Sample Collection				Metadata Type (Field Name)
3	isolate metadata		Data which describes an isolate.		Required Subfields
4		isolate ID	Identifier of the specific isolate.	SARS-CoV-2/human/USA/CA-CDP-H-001/2020; MPXV_USA_2022_MA001	Definition
5		isolation date / isolate collection date	The date on which the isolate was isolated from a sample.	2023-08; 2023-08-23; April 5th, 2021	Example Subfields
6		isolate received date	The date on which the isolate was received by the laboratory.	2023-08; 2023-08-23; April 5th, 2021	Values
7		isolated by	The name of the agency, organization or institution with which the individual who performed the isolation procedure is affiliated.	Public Health Agency of Canada (PHAC); Agriculture and Agri-Food Canada (AAFC); LifeLabs; Dynacare; Mount Sinai Hospital	Motivation
8	organism		Taxonomic name of the organism.	Severe acute respiratory syndrome coronavirus 2; Acinetobacter baumannii; Aeromonas; Enterobacter cloacae	
9	purpose of sampling		The reason that the sample was collected.	Diagnostic testing; Survey study; Protocol testing; Surveillance; Environmental testing; Cluster/Outbreak investigation; Field experiment	
10	geographical location (sample)		Data which describes the location of origin of the sample.		
11		geo_loc name (country)	The country where the sample was collected.	South Africa; Australia; Bahamas; Canada; United Kingdom; United States of America	
12		geo_loc name (state/province/territory)	The state/province/territory where the sample was collected.	Western Cape; British Columbia; Provence; California	
13		geo_loc name (county/region)	The county/region of origin of the sample.	Derbyshire; Orange County; Queens; Harris	

"Metadata Legend" with formatted cells, cell contents imported.

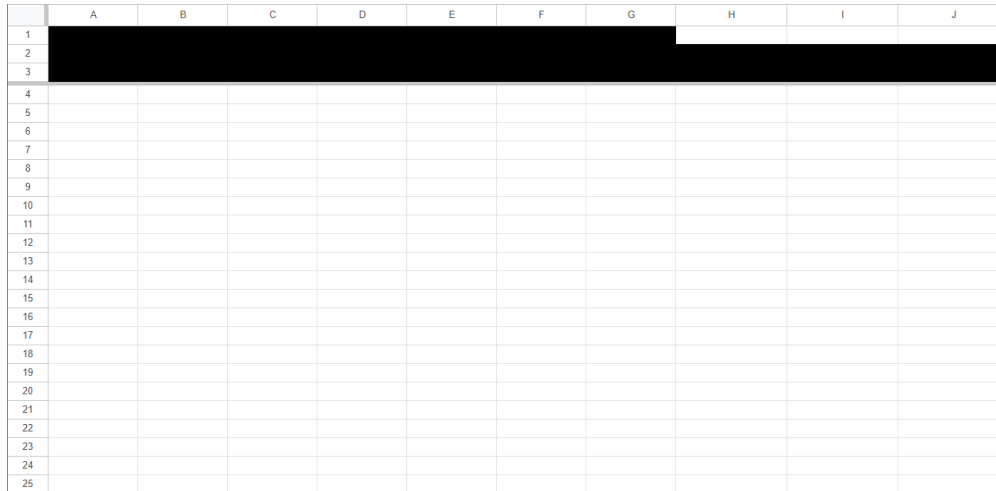
23.3 Protect non-editable import cells / ranges.

A1:H108	Sort range				
	A	B	C	D	E
1	Metadata Type (Field Name)	Subfield (Required in Bold)	Description / Definition	Example	Header
2	Sample Collection				Metadata Type (Field Name)
3	isolate metadata		Data which describes an isolate.		Required Subfields
4		isolate ID	Identifier of the specific isolate.	SARS-CoV-2/human/USA/CA-CDP-H-001/2020; MPXV_USA_2022_MA001	Definition
5		isolation date / isolate collection date	The date on which the isolate was isolated from a sample.	2023-08; 2023-08-23; April 5th, 2021	Example Subfields
6		isolate received date	The date on which the isolate was received by the laboratory.	2023-08; 2023-08-23; April 5th, 2021	Values
7		isolated by	The name of the agency, organization or institution with which the individual who performed the isolation procedure is affiliated.	Public Health Agency of Canada (PHAC); Agriculture and Agri-Food Canada (AAFC); LifeLabs; Dynacare; Mount Sinai Hospital	Motivation
8	organism		Taxonomic name of the organism.	Severe acute respiratory syndrome coronavirus 2; Acinetobacter baumannii; Aeromonas; Enterobacter cloacae	
9	purpose of sampling		The reason that the sample was collected.	Diagnostic testing; Survey study; Protocol testing; Surveillance; Environmental testing; Cluster/Outbreak investigation; Field experiment	
10	geographical location (sample)		Data which describes the location of origin of the sample.		
11		geo_loc name (country)	The country where the sample was collected.	South Africa; Australia; Bahamas; Canada; United Kingdom; United States of America	
12		geo_loc name (state/province/territory)	The state/province/territory where the sample was collected.	Western Cape; British Columbia; Provence; California	
13		geo_loc name (county/region)	The county/region of origin of the sample.	Derbyshire; Orange County; Queens; Harris	

"Metadata Legend" with formatted cells, cell contents imported and protected from editing.

24 Import "OVERVIEW TABLE" Template:

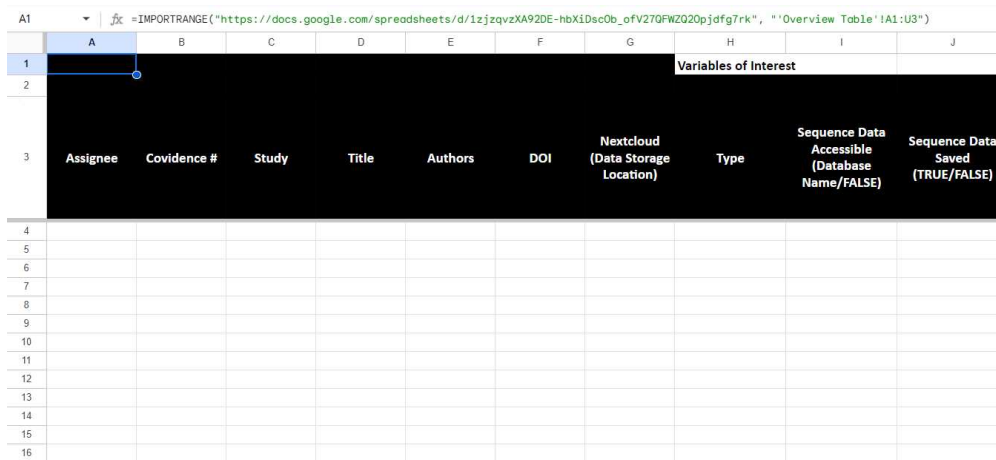
- 24.1 **Apply formatting** (Optional): Copy the formatted "Reference" version and paste it (with formatting) into a new tab and then delete all cell values (do not clear the formatting).



"Overview Table" with formatted cells, cell contents deleted.

- 24.2 **Import columns and/or rows** (applicable range, don't forget your version info).
EXAMPLE:

```
=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjzqvzXA92DE-hbXiDsc0b_ofV27QFWZQ20pjdfg7rk", "'Overview Table'!A1:U3")
```



"Overview Table" with formatted cells, cell contents imported.

- 24.3 **Protect non-editable import cells / ranges** - in this tab it will be your header rows because you want users to input data below them.

A1:U3

	A	B	C	D	E	F	G	H	I	J
1	Variables of Interest									
2										
3	Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type	Sequence Data Accessible (Database Name/FALSE)	Sequence Data Saved (TRUE/FALSE)
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										

"Overview Table" with header rows selected/highlighted to receive editing protections.

24.4 Create any data input functionalities you want and apply to the entire column. EXAMPLE:

H4

A

1

2

3

Assignee

Cov

4

5

6

7

8

9

10

11

12

13

14

15

16

Columns

Sheet

Tables

Chart

Pivot table

Image

Drawing

Function

Link

Checkbox

Dropdown

Emoji

Smart chips

Shift+F11

Ctrl+K

F

G

H

I

J

Variables of Interest

ors

DOI

Nextcloud
(Data Storage
Location)

Type

Sequence Data
Accessible
(Database
Name/FALSE)

Sequence Data
Saved
(TRUE/FALSE)

Adding custom dropdown menu for enumeration data entry.

	A	B	C	D	E	F	G	H	I	J
1								Variables of Interest		
2										
3	Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type	Sequence Data Accessible (Database Name/FALSE)	Sequence Data Saved (TRUE/FALSE)
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										

The resulting dropdown menu.

	A	B	C	D	E	F	G	H	I	J
1								Variables of Interest		
2										
3	Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type	Sequence Data Accessible (Database Name/FALSE)	Sequence Data Saved (TRUE/FALSE)
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										

Copying the dropdown menu, then using "CTRL/CMD+Shift+down" to select the remainder of the column before *pasting*.

24.5 Import "METADATA TABLE" Template:

- 25 **Apply formatting** (Optional): Copy the formatted "Reference" version and paste it (with formatting) into a new tab and then delete all cell values (do not clear the formatting).

	A	B	C	D	E	F	G	H
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								

"Metadata Table" with formatted cells, cell contents deleted.

25.1 Import columns and/or rows (applicable range, don't forget your version info). EXAMPLE:

```
=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjqvzXA92DE-hbXiDsc0b_ofV27QFWZQ20pjdfg7rk", "'Metadata Table'!A1:D83")
```

A1	=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjqvzXA92DE-hbXiDsc0b_ofV27QFWZQ20pjdfg7rk", "'Metadata Table'!A1:D83")							
	A	B	C	D	E	F	G	H
1	Metadata Type (Field Name)							
2	Version 1	Required Subfield	Definition	Example Subfield				
3	Sample Collection							
4	isolate metadata							
5		Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	Data which describes an isolate	isolation date / isolate collection date isolate received date isolated by isolate database accession ID				
6	organism	isolate ID	Identifier of the specific isolate					
7	purpose of sampling	Quote value if example found	Taxonomic name of the organism					
8	geographical location (sample)	Quote value if example found	The reason that the sample was collected.					
9		Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	A data field which describes the location of origin of the sample.	geo_loc name (county/region) geo_loc name (city) geo_loc name (site) geo_loc latitude geo_loc longitude food product origin geo_loc name (country)				
10		geo_loc name (country)	The country where the sample was collected.					
11	sample collected by	geo_loc name (state/province/territory)	The state/province/territory where the sample was collected.					

"Metadata Table" with formatted cells, cell contents imported.

25.2 Add header row of the data input columns. You can do this manually or make the reference work so you can import them (this may result in your user seeing more columns than they need for data entry).



E2:CP2 ▾ *fx* Values

	A	B	C	D	E	F	G	H
1	Metadata Type (Field Name)							
2		Required Subfield	Definition	Example Subfield	Values	Motivation	Values	Motivation
3	Sample Collection							
4	isolate metadata							
5		Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	Data which describes an isolate.	isolation date / isolate collection date isolate received date isolate database accession ID				
6	organism	isolate ID	Identifier of the specific isolate.					
7	purpose of sampling	Quote value if example found	Taxonomic name of the organism.					
8	geographical location (sample)	Quote value if example found	The reason that the sample was collected.					
9		Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	A data field which describes the location of origin of the sample.	geo_loc name (county/region) geo_loc name (city) geo_loc name (site) geo_loc latitude geo_loc longitude food product origin geo_loc name (country)				
10		geo_loc name (country)	The country where the sample was collected.					
11	sample collected by	geo_loc name (state/province/territory)	The state/province/territory where the sample was collected.					

Adding "Value" and "Motivation" column labels.

25.3 Protect non-editable import cells / ranges - the "header columns" and optionally the data input header row.

A1:D83 ▾ *fx* =IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjqvzXA920E-hbX1Dsc0e_ofv27QfWZ020pjdfig7rk", "Metadata Table!A1:D83")

	A	B	C	D	E	F	G	H
1	Metadata Type (Field Name)							
2		Required Subfield	Definition	Example Subfield	Values	Motivation	Values	Motivation
3	Sample Collection							
4	isolate metadata							
5		Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	Data which describes an isolate.	isolation date / isolate collection date isolate received date isolate database accession ID				
6	organism	isolate ID	Identifier of the specific isolate.					
7	purpose of sampling	Quote value if example found	Taxonomic name of the organism.					
8	geographical location (sample)	Quote value if example found	The reason that the sample was collected.					
9		Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	A data field which describes the location of origin of the sample.	geo_loc name (county/region) geo_loc name (city) geo_loc name (site) geo_loc latitude geo_loc longitude food product origin geo_loc name (country)				
10		geo_loc name (country)	The country where the sample was collected.					
11	sample collected by	geo_loc name (state/province/territory)	The state/province/territory where the sample was collected.					

"Overview Table" with header and support columns to receive editing protections.

Pilot Preparation

- 26 **Decide on how many papers you want the pilot to cover and which papers.** If possible, prepare at least five pilot papers for review - one that will be an example, two for the pilot, and two more for additional practice for everyone or at least those who struggled with the first pilot. Choose papers that get a mix of study types (e.g. one that focuses on a disease outbreak, one that focuses on surveillance, at least one with supplementary files, at least one with accession data, etc.).



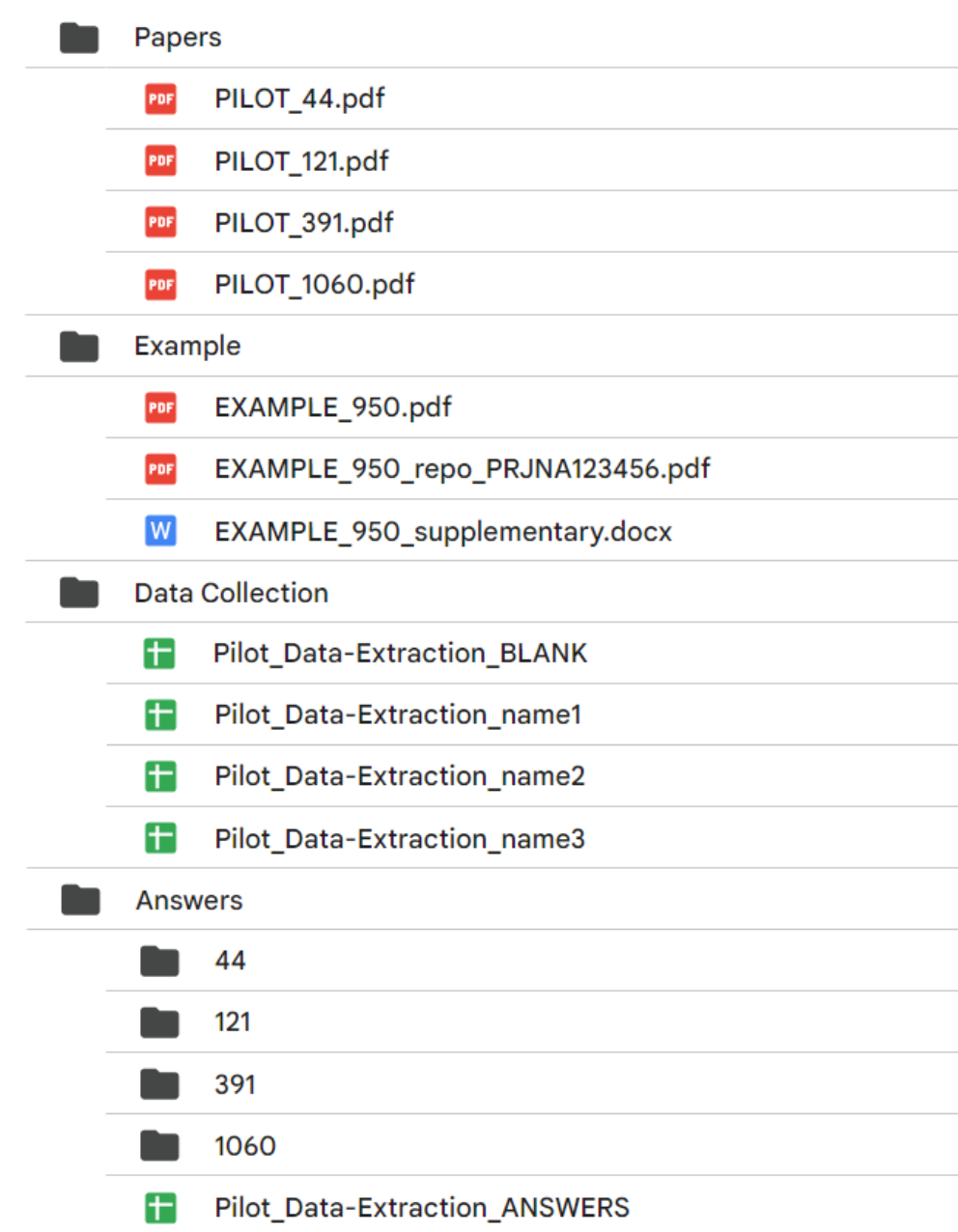
Note

IMPORTANT: Select papers that are under your assignment so that (A) your version is the one in the final dataset and (B) other data extractors don't end up accidentally repeating your work. If you would like to use a paper assigned to someone else just rebalance the assignments between you and that individual by giving them one of yours.

27 **Prepare your files.** We used Google Drive because this is compatible with the data collection from Google Sheets; you could also consider using Nextcloud. Here is a possible folder layout:

- **Papers** - The papers users will practice extracting data from.
- **Example** - Example results that users can reference as they do their data extraction.
- **Data Collection** - Copies of the data collection form assigned to the different collectors, each data collector should only have access to their own worksheet.
- **Answers** - Where the answers for the pilot will be, including the additional files that should have been identified and undergone extraction.

EXAMPLE

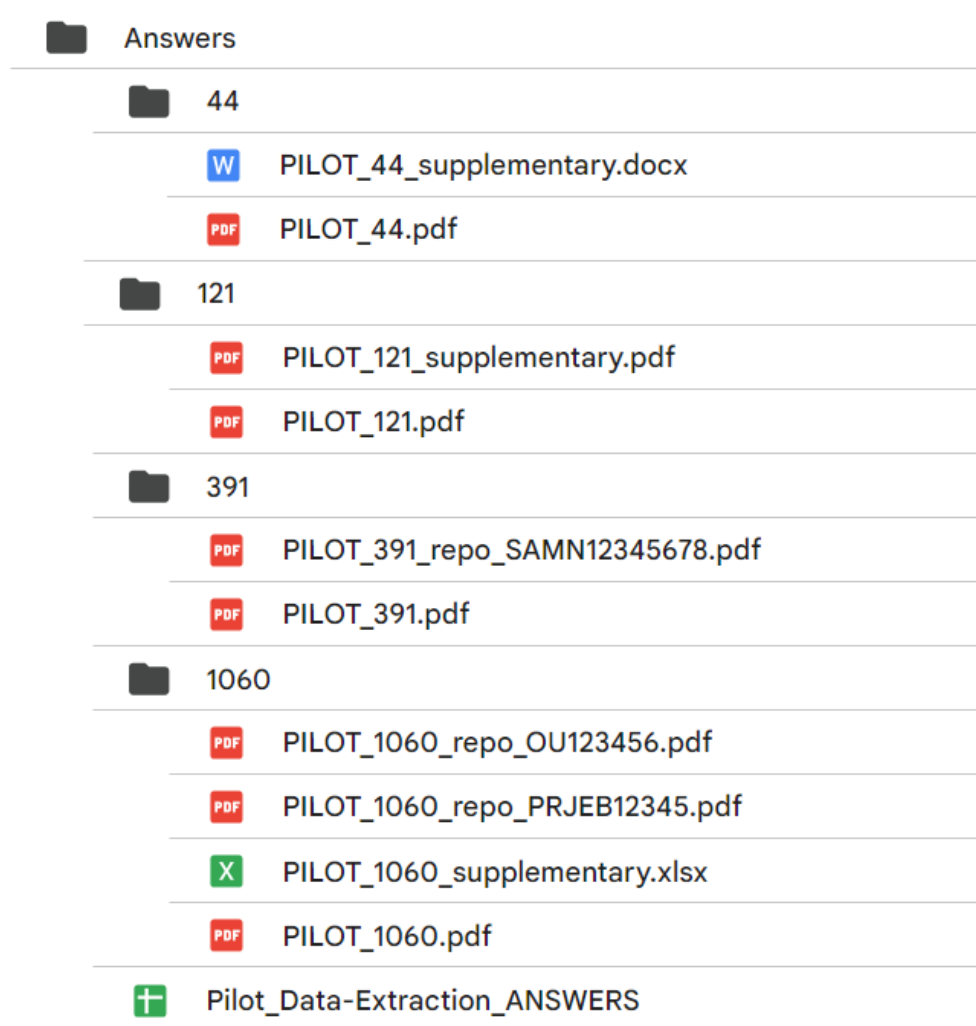


What the final pilot workspace may look like (with "Answer" subfolders collapsed).

- 27.1 **Download your papers, rename them with their assigned "Study/Covidence #".**
- 27.2 **Download all the supplementary files you can find into "Answers".** Be sure to check inside the papers for data availability statements and accessions - when you find any, navigate to a single sample accession and save it as a PDF file. If you are dealing with a small dataset you can choose to check every sample accession, but that can be repetitive and inefficient when there are hundreds or even thousands of samples.

Note

LIMITATION: It is possible that different samples have different contextual metadata available; this will be lost by the selection of only one random sample.



An example "Answers" folder with expanded subfolders.

- 27.3 **Rename the supplementary files so that they all reference the associated "Study/Covidence #" and then upload them to "Study/Covidence #" folders within the "Answers" folder.**
- 27.4 **Make a copy of each paper from "Papers" in their respective "Answers" subfolder.** These are the copies you will use as you prepare the answer key - highlighting text you decide to extract and adding comments as you so choose. This way if data extractors



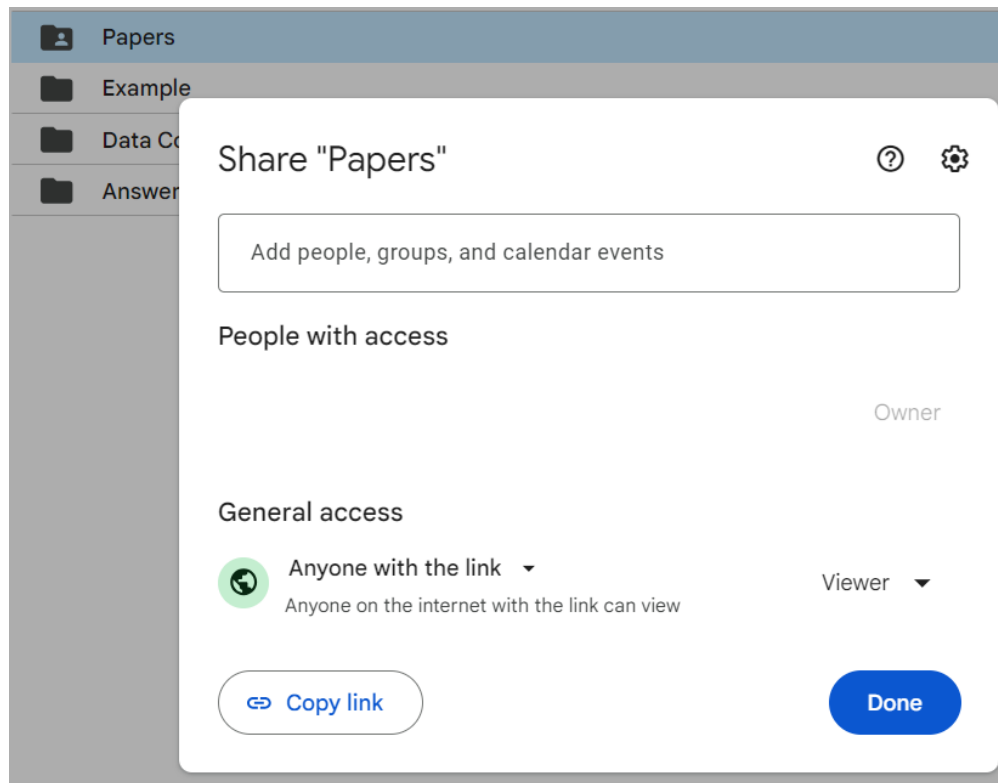
are confused by your answers (e.g. finding a "TRUE" where they had indicated "FALSE") they can look at the paper, find the highlighted text, and see the context of your extraction.

- 27.5 **Start a protocol document of some sort.** This can be a Google Doc, protocols.io, or in the reference README (so that it automatically updates the collection templates). You may find you want to adjust the guidance provided in this protocol to be specific to your setup as you go through generating the answers. If you aren't putting this directly in the README file, now is a good time to **add a link to this resource in your reference README.**
- 27.6 **Create a copy of version 1 of your data collection template (not the reference) and put it in the "Answers" folder.** We will reference this file as "Pilot_Data-Extraction_ANSWERS" moving forward. Do not make copies for participants yet.
- 27.7 Update the **"Overview Table"** in **"Pilot_Data-Extraction_ANSWERS"** so that only the papers relevant to the pilot are listed. Also remove the **"Data Storage Location"** information, these are going to be reserved for the final data extraction.

Variables of Interest									
Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type	Sequence Data Accessible (Database Name/FALSE)	Sequence Data Saved (TRUE/FALSE)
EXAMPLE	950	AlKalamouni 202	Genomic surveill	AlKalamouni H; /10.1186/s12920					
PILOT	391	Barkley 2016	Whole-Genome	Barkley JS; Gosci 10.3201/eid2206					
PILOT	1060	Boumasmoud 20	Genomic Surveil	Boumasmoud M 10.1128/mbio.0					
PILOT	44	Berggreen 2023	Lessons learned:	Berggreen H; LÃ 10.1016/j.jhin.20					
PILOT	121	Bainomugisa 20	Multi-clonal evol	Bainomugisa A; 110.1099/mgen.0					

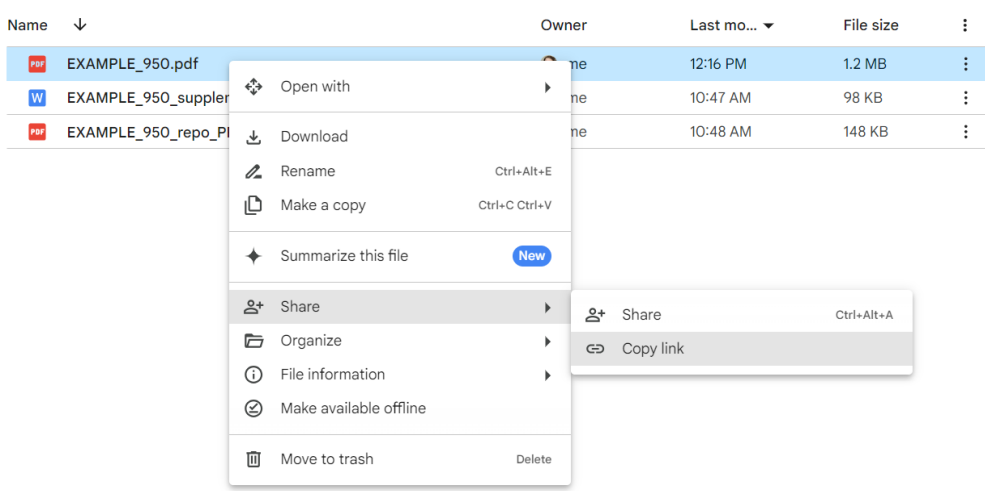
"Overview Table" show basic information for the example and pilot papers.

- 27.8 Make sure the **"Papers" folder's "General access"** is set so that **"Anyone with the link"** is a **"Viewer"** so that participants can access the files to view/download but not edit/delete.



Permissions settings for a Google Drive document.

27.9 Update the "Overview Table" in "Pilot_Data-Extraction_ANSWERS" with paper URLs under "Data Storage Location".



Getting the study URL.

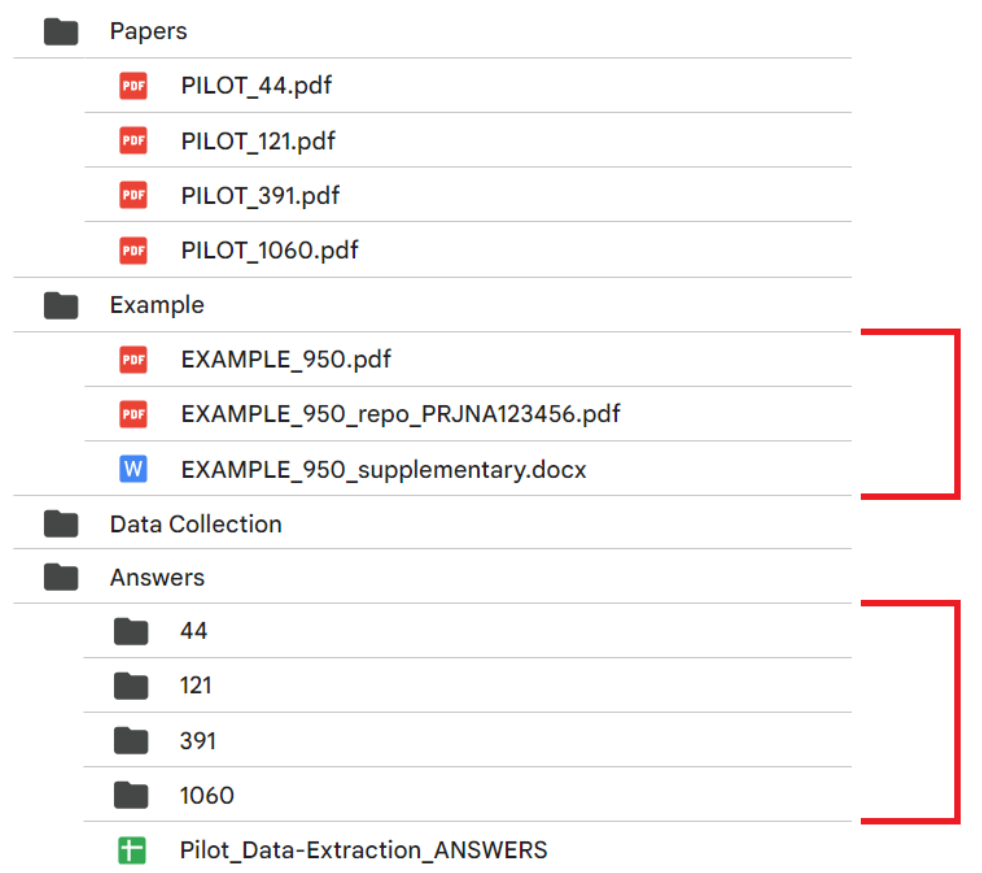
Variables of Interest									
Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type	Sequence Data Accessible (Database Name/FALSE)	Sequence Data Saved (TRUE/FALSE)
EXAMPLE	950	AlKalamouni 202	Genomic surveill	AlKalamouni H; J	10.1186/s12920	https://drive.google.com/...			
PILOT	391	Barkley 2016	Whole-Genome	Barkley JS; Gosci	10.3201/eid2206	https://drive.google.com/...			
PILOT	1060	Boumasmoud 20	Genomic Surveill	Boumasmoud M	10.1128/mbio.0	https://drive.google.com/...			
PILOT	44	Berggreen 2023	Lessons learned	Berggreen H; LÃ	10.1016/j.jhin.20	https://drive.google.com/...			
PILOT	121	Bainomugisa 20	Multi-clonal evol	Bainomugisa A; I	10.1099/mgen.0	https://drive.google.com/...			

Listing URLs under "Data Storage Location" in the "Overview Table".

- 28 **Prepare the answers by going through the "Data Extraction (Pilot)" steps yourself for all of the pilot studies ("Example" and "Answers" papers, as well as any additional supporting materials).** You want to leave the "Papers" folder documents untouched and without highlighted text because these are what participants download for their attempts.

Note

While the "Data Extraction (Pilot)" procedure says highlighting extracted text in sources is optional, it is not optional for you since you are preparing the resources that participants are using to understand your decisions and double check their work. When developing answer keys as such, it is advisable to highlight the sources of extracted text.



The resources you will be highlighting.

- 28.1 **Make sure "Pilot_Data-Extraction_ANSWERS" is under "Restricted" access** in case someone gets curious and follows the import link from their template, accidentally exposing themselves to the answer key before they have attempted the pilot data extraction themselves.
- 29 **Prepare participant collection templates.**
 - 29.1 Make a copy of "**Pilot_Data-Extraction_ANSWERS**", move it to "**Data Collection**" and rename it "**Pilot_Data-Extraction_BLANK**".
 - 29.2 Make sure the "**README**" editable fields in "**Pilot_Data-Extraction_BLANK**" are replaced with placeholders and not your information.



B2 | <YYYY-MM-DD> to <YYYY-MM-DD>

A	B	C
1	Data Extractor Name:	<First Name> <Last Name>
2	Date Range of Extraction:	<YYYY-MM-DD> to <YYYY-MM-DD>
3		
4	General:	This is your personal worksheet for the scoping review full-text metadata extraction. All headers are imported from a master reference sheet that will be updated as necessary to ensure everyone has the most up-to-date and comprehensive information available. It is recommended you familiarize yourself with and frequently reference the "Legend" tabs, as they are your guide to filling in the tables. All data extraction tables have a aforementioned "Legend" that describes headers, and provides value examples/descriptions of what those fields are for. Tables are designed so that they import reference information from a reference sheet; when the reference sheet is

"README" tab with placeholder text in the fields participants can edit.

- 29.3 In **"Pilot_Data-Extraction_BLANK"**, delete your extracted answers for the **"PILOT"** papers in the **"Overview Table"** and **"Metadata Table"**.

Variables of Interest									
Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type	Sequence Data Accessible (Database Name/FALSE)	Sequence Data Saved (TRUE/FALSE)
EXAMPLE	950	AlKalamouni 202	Genomic surveill	AlKalamouni H; /	10.1186/s12920	https://drive.google.com/...	Surveillance	FALSE	TRUE
PILOT	391	Barkley 2016	Whole-Genome	Barkley JS; Gosci	10.3201/eid2201	https://drive.google.com/...			
PILOT	1060	Boumasmoud 20	Genomic Surveil	Boumasmoud M	10.1128/mbio.0	https://drive.google.com/...			
PILOT	44	Berggreen 2023	Lessons learned: Berggreen H; LÃ	10.1016/j.jhin.20		https://drive.google.com/...			
PILOT	121	Bainomugisa 20	Multi-clonal evol	Bainomugisa A; I	10.1099/mgen.0	https://drive.google.com/...			

"Overview Table" with the data collection fields emptied for "PILOT" rows.

- 29.4 In **"Pilot_Data-Extraction_BLANK"**, replace your **"Overview Table"** **"EXAMPLE"** row with an imported version that takes the data from **"Pilot_Data-Extraction_BLANK"**. Don't forget to protect the import range from edits.

EXAMPLE FUNCTION:

```
=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjzqvzXA92DE-hbXiDscOb_ofV27QFWZQ20pjdfg7rk", "'Overview Table'!A4:U4")
```



A4 =IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjzqvzXA92DE-hbXiDsc0b_ofV27QFWZQ20pjdfg7rk", "'Overview Table'!A4:U4")

	A	B	C	D	E	F	G	H	I	J
1								Variables of Interest		
2										
3										
	Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type	Sequence Data Accessible (Database Name/FALSE)	Sequence Data Saved (TRUE/FALSE)
4	EXAMPLE	950	AlKalamouni 2023	Genomic surveill	AlKalamouni H; #10.1186/s12920	https://drive.google.com/...	Surveillance		FALSE	TRUE
5	PILOT	391	Barkley 2016	Whole-Genome	Barkley JS; Gosci 10.3201/eid2201	https://drive.google.com/...				
6	PILOT	1060	Boumasmoud 2023	Genomic Surveill	Boumasmoud M 10.1128/mbio.0	https://drive.google.com/...				
7	PILOT	44	Berggreen 2023	Lessons learned	Berggreen H; LÃ 10.1016/j.jhin.20	https://drive.google.com/...				
8	PILOT	121	Bainomugisa 2023	Multi-clonal evol	Bainomugisa A; 110.1099/mgen.0	https://drive.google.com/...				

"Overview Table" with the "EXAMPLE" row replaced by an imported version.

Note

Importing rows is also how you will eventually provide answers to participants. I.e. you will add import lines to their specific worksheets. People work at their own pace, so if you give everyone the answer key at a specific time some may not be ready for it. This is also an easy way for participants to check against their answers, and you can limit which answers they see (e.g. only show the answers for the first two pilot papers in case they later decided to do the final set of pilot papers).

- 29.5 In "**Pilot_Data-Extraction_BLANK**", replace your "**Metadata Table**" "**EXAMPLE**" columns with an imported version that takes the data from "**Pilot_Data-Extraction_BLANK**". Don't forget to protect the import range from edits.

EXAMPLE FUNCTION:

```
=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1zjzqvzXA92DE-hbXiDsc0b_ofV27QFWZQ20pjdfg7rk", "'Metadata Table'!E1:F83")
```

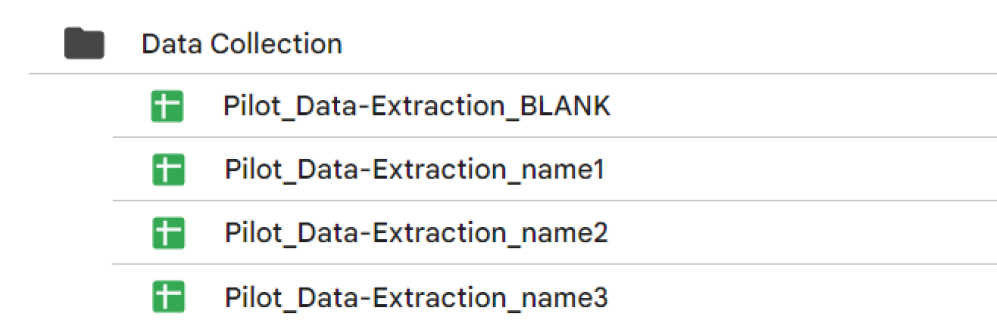


E1 [=IMPORTRANGE\("https://docs.google.com/spreadsheets/d/1zjqvzXA92DE-hbX1Dsc0b_ofV27QfWZQ2Opjdfg7rk", "Metadata Table"!E1:F83\)](#)

	A	B	C	D	E	F	G	H
1	Metadata Type (Field Name)				950 (Example)	950 (Example)	44	44
2	Version 1	Required Subfield	Definition	Example Subfield	Values	Motivation		
3	Sample Collection							
4	Isolate metadata							
5		Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	Data which describes an isolate.	isolation date / isolate collection date isolate received date isolate database accession ID	FALSE	FALSE		
6	organism		Identifier of the specific isolate.		FALSE	FALSE		
					SARS-CoV-2: Omicron BA.1.1 sub-lineage, Omicron BA.1 sub-lineage, Omicron BA.2 sub-lineage, Delta B.1.617.2 sub-lineage			
7	purpose of sampling	Quote value if example found.	Taxonomic name of the organism.		Surveillance	FALSE		
8	geographical location (sample)	Quote value if example found.	The reason that the sample was collected.		TRUE	TRUE		
9		Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	A data field which describes the location of origin of the sample.	geo_loc name (county/region) geo_loc name (city) geo_loc name (site) geo_loc latitude geo_loc longitude food product origin geo_loc name (country)				
10		geo_loc name (country)	The country where the sample was collected.		Lebanon	FALSE		
					South Lebanon, Mount Lebanon	A key limitation of our study is that it did not include samples from all regions in Lebanon and thus is not fully representative of the situation in Lebanon.		
11	sample collected by	geo_loc name (state/province/territory)	The state/province/territory where the sample was collected.		TRUE	FALSE		
		Is TRUE if required subfield found. Quote value if example subfield. Otherwise FALSE	Data which describes the information regarding who performed the sample collection and how.		American University of Beirut (AUB) Medical Center, Rasoul Al Azam Hospital, Belle Vue Hospital			

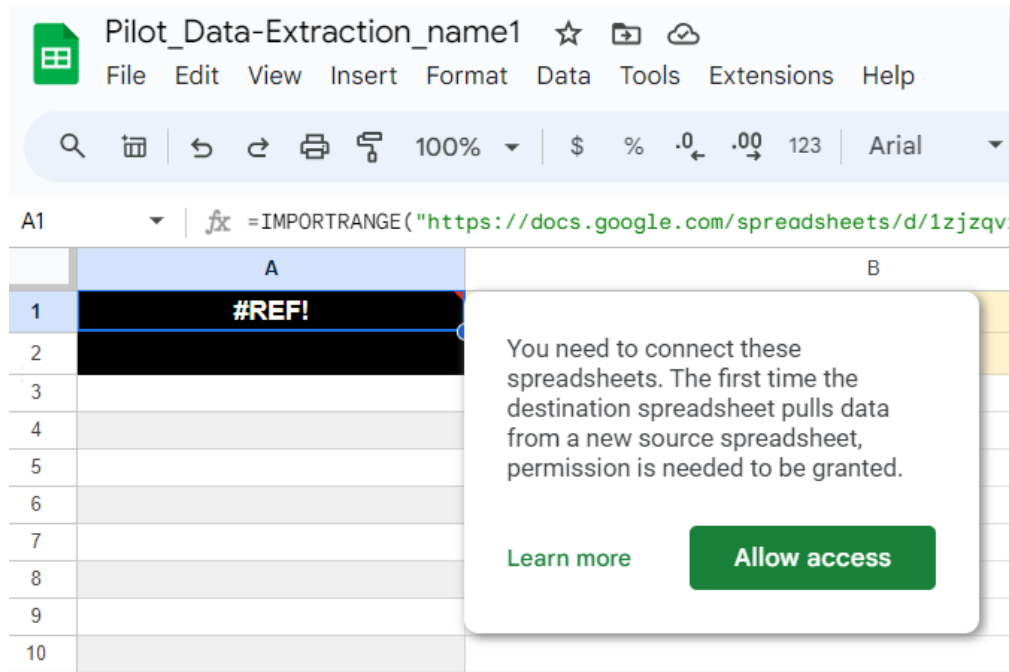
"Metadata Table" with the "EXAMPLE" columns replaced by an imported version.

- 29.6 Make copies of **"Pilot_Data-Extraction_BLANK"** in the **"Data Collection"** folder and label with the respective participant names. Do not overwrite **"Pilot_Data-Extraction_BLANK"** in case you end up with more participants joining at a later time.



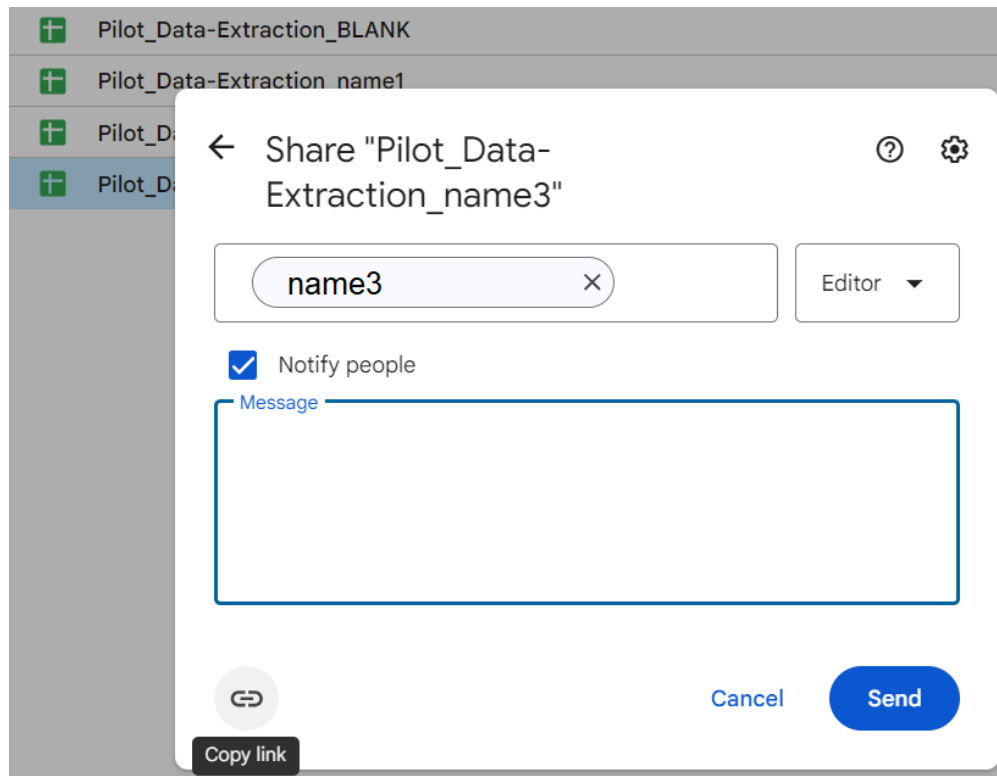
"Data Collection" templates for individual participants and a backup.

- 29.7 Open all **"Pilot_Data-Extraction_name"** collection templates and confirm all the import functions have access to the file they are importing from.



"README" import function showing a "#REF!" error because it has not been allowed access to the spreadsheet it is trying to import from.

- 29.8 Do a **final review of the protocol** and polish up anything you can think of.
- 29.9 Set-up permissions so participants have "**Editor**" access to their individual "**Pilot_Data-Extraction_name**" worksheet. You may also want to copy the "**Share**" link and send it to them directly using your normal means of communication so that they don't accidentally miss the notification.

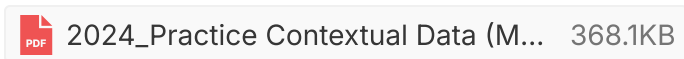


Permissions settings for a Google Drive document - giving editor access to an individual instead of making the entire document editable via a URL.

Data Extraction (Pilot)

30 Practice Contextual Data (Metadata) Extraction Scenario (Optional)

Run through this practice extraction scenario designed by Dr. Emma Griffiths. It will give you a sense of how you will be extracting data for this protocol.



- 31 Locate your individual **"Pilot_Data-Extraction_name"** collection form and confirm you have edit permissions.
- 32 Review the **"README"** and the **pilot data extraction protocol**.
- 33 Go to the **"Overview Table"** and download the **"PILOT"** papers using the URLs listed under **"Data Storage Location"**.

- 34 Find all associated **supplementary data** using information within the paper and/or the **DOI** listed in the **"Overview Table"**. Look for data availability statements and reference to databases and database accessions. Download and label with the **"Study/Covidence #"** of the associated paper.
- 35 Fill in the **"Overview Table"** for the study according to the guidance in your **protocol** and the **"Overview Legend"**. We recommend highlighting the information within the files as you extract to make it easier to review later.

Note

Depending on your preference, you can fill in the "Overview Table" for all studies and then do the "Metadata Table" for all studies, or you can fill in the "Overview Table" for one study and then immediately do the "Metadata Table" for the same study. We recommend you do the latter for the pilot, so that you keep your head in one "study" at a time, but when you get the hang of the process it tends to be more efficient to do the former.

- 36 Fill in the **"Metadata Table"** for the study according to the guidance in your **protocol** and the **"Metadata Legend"**. We recommend highlighting the information within the files as you extract to make it easier to review later.
- 37 Review both the **"Overview Table"** and **"Metadata Table"** for completeness. **Blank cells are not the same** as **"FALSE"**. If you notice blanks you should re-review your sources to see if you can find a value that specifically indicates the presence of a variable. If you do not find any, you can then indicate **"FALSE"**.
- 38 **Request Answers / Provide Answers.** The methods manager will add in import lines for their solutions into to your **"Pilot_Data-Extraction_name"** for the studies you worked on.
- 38.1 **Check your answers against the key.**
- Did you find all the **supplementary files**?
 - Did you find the **accession** associated data? It doesn't have to be the exact same sample.
 - Did you have **extracted metadata** where the key had extracted metadata? I.e. did you have "TRUE" values where the answer key had "TRUE" values, even if you didn't extract the same metadata value?
 - Did you identify similar **objectives**?
- 39 **Post-Pilot Group Discussion.**
- Were there any problems with the answer key?
 - Was there any confusion on how to extract data for specific fields?



- Can improvements/clarifications be made to the Legends or Guidance?
- Are there any variables of interest that should be changed?

40 **Update reference resources and extend pilot as necessary.**

Give users who struggled an additional two papers to review. If they had already gone through all the prepared answers, prepare more. Doing more will not be a waste of time as they will still be used as part of data collection and it is important you data collectors feel self-assured in this process before the final data extraction.

Note

REMINDER: Keep updating your versioning as you make changes.

Final Preparation

- 41 **Prepare the data collection forms.** Once again using Google Drive and Sheets. Make a copy of your most up-to-date data extraction form ("**Final_Data-Extraction_BLANK**") and move it to a new "**Data Collection (Final)**" folder.

Note

CHECK: The reference template used to make "**Final_Data-Extraction_BLANK**" should be updated and refined from the pilot.

- 42 Make sure the "**README**" editable fields in "**Final_Data-Extraction_BLANK**" are replaced with placeholders and not your information.
- 43 Make copies of "**Final_Data-Extraction_BLANK**" and label with the respective participant names. Do not overwrite "**Final_Data-Extraction_BLANK**" in case you end up with more participants joining at a later time.



 Final_Data-Extraction_BLANK

 Final_Data-Extraction_name1

 Final_Data-Extraction_name2

 Final_Data-Extraction_name3

Data collection forms with a blank version being kept as backup.

- 44 Go to your **Main/Master "Overview Table"** where the complete list of papers are and make sure it is sorted by **"Assignee"**.

Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type	Sequence Data Accessible (Database Name/FALSE)
Extractor 3	30	Lachmann 2022	Invasive listeriosis	Lachmann R; Hal	10.1080/222217	https://cloud.lab		
Extractor 3	136	Simon 2019	Whole Genome	Simon B; Pichon	10.3390/v11020	https://cloud.lab		
Extractor 3	146	Enkelmann 2020	Disentangling ou	Enkelmann J; voi	10.1017/S09502	https://cloud.lab		
Extractor 1	165	Houlihan 2018	Use of Whole-Ge	Houlihan CF; Fra	10.1093/infdis/ji	https://cloud.lab		
Extractor 2	223	Magalhães 202	Combining Stanc	Magalhães B; V	10.3389/fpubh.2	https://cloud.lab		
Extractor 3	224	Sui 2018	Whole genome s	Sui W; Zhou H; D	10.1186/s13756	https://cloud.lab		
Extractor 3	228	Comán 2022	The MtZ Strain: I	Comán J; Madac	10.3389/fcimb.2	https://cloud.lab		
Extractor 2	249	Underwood 201	Public health val	Underwood AP; 1	10.1128/JCM.01	https://cloud.lab		
Extractor 2	257	Casali 2016	Whole Genome	Casali N; Broda /	10.1371/journal	https://cloud.lab		
Extractor 3	280	Elson 2019	Utility of Whole	Elson R; Awofisa	10.4315/0362-0	https://cloud.lab		
Extractor 3	295	Toleman 2016	Systematic Surve	Toleman MS; Rei	10.1093/infdis/ji	https://cloud.lab		
Extractor 1	311	Wendel 2018	One-year molecu	Wendel AF; Male	10.1186/s13756	https://cloud.lab		
Extractor 2	330	Weßels 2018	Emergence and i	Weßels C; Stror	10.1093/jac/dky	https://cloud.lab		
Extractor 3	367	Britto 2020	Persistent circula	Britto CD; Dyson	10.1093/jac/dkz	https://cloud.lab		
Extractor 2	397	CunhaMDP 2015	Origin of the SÄ	Cunha MDP; Du	10.1038/s41598	https://cloud.lab		
Extractor 1	398	Smith 2019	Outbreak of List	Smith AM; Tau N	10.1089/fpd.201	https://cloud.lab		
Extractor 2	399	PereiraGusmãE	Return of the fo	Pereira GusmãE	10.1080/222217	https://cloud.lab		
Extractor 3	413	Morganti 2018	Rise and fall of	Morganti M; Bol	10.2807/1560-7	https://cloud.lab		
Extractor 3	431	vanderZwaluw 2	Molecular chara	van der Zwaluw	10.1016/j.cmi.20	https://cloud.lab		

Main/Master "Overview Table" sorted by "Study/Covidence #".



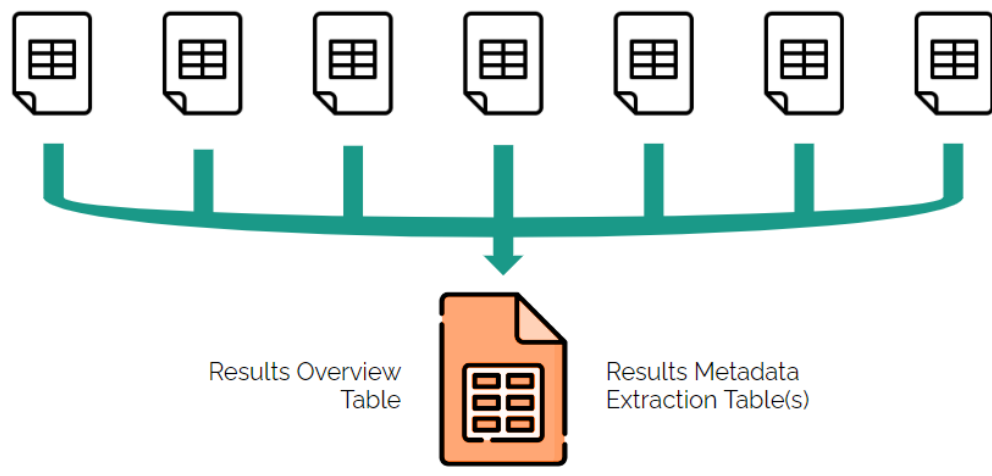
Assignee	Covidence #	Study	Title	Authors	DOI	Nextcloud (Data Storage Location)	Type	Sequence Data Accessible (Database Name/FALSE)
Extractor 1	165	Houlihan 2018	Use of Whole-Genome Sequencing to Investigate the Role of the	Houlihan CF; Frazer	10.1093/infdis/ji	https://cloud.lab		
Extractor 1	311	Wendel 2018	One-year molecul	Wendel AF; Malik	10.1186/s13756	https://cloud.lab		
Extractor 1	398	Smith 2019	Outbreak of List	Smith AM; Tau N	10.1089/fpd.201	https://cloud.lab		
Extractor 1	494	Senn 2016	The Stealthy Sup	Senn L; Clerc O;	10.1128/mBio.0	https://cloud.lab		
Extractor 1	581	Lazaris 2017	Novel multiresist	Lazaris A; Coleman	10.1093/jac/dkx	https://cloud.lab		
Extractor 1	610	Chen 2019	Molecular Epide	Chen FJ; Huang	10.1128/AAC.02	https://cloud.lab		
Extractor 1	615	Marimuthu 2017	Clinical and Mol	Marimuthu K; Ve	10.1093/cid/cix1	https://cloud.lab		
Extractor 1	691	Davidson 2022	Genomic Analysi	Davidson RM; Ni	10.1128/JCM.01	https://cloud.lab		
Extractor 1	696	Chaves 2017	High mortality o	Chaves L; Tomich	10.1099/jmm.0	https://cloud.lab		
Extractor 1	734	GelbÃrovÃi 20	An outbreak of li	GelbÃrovÃi T;	10.1017/S09502	https://cloud.lab		
Extractor 1	827	Zhu 2021	Longitudinal Ger	Zhu C; Li C; Lai C	10.1016/j.ijantin	https://cloud.lab		
Extractor 1	948	Avendano 2022	SARS-CoV-2 Vari	Avendano C; Lili	10.1001/jamane	https://cloud.lab		
Extractor 1	987	Findlay 2017	Characterization	Findlay J; Hopkin	10.1093/jac/dkx	https://cloud.lab		
Extractor 1	997	Ghodousi 2020	Genomic analysi	Ghodousi A; Bor	10.1371/journal	https://cloud.lab		
Extractor 1	998	FrancÃs-Cuest	Is there a wides	FrancÃs-Cuest	10.1016/j.jhin.2	https://cloud.lab		
Extractor 1	1014	deLemosAPS 20	Genomic surveill	de Lemos APS; S	10.1371/journal	https://cloud.lab		
Extractor 1	1025	Liu 2022	Clonal Spread of	Liu X; Wang K; Cl	10.1128/spectru	https://cloud.lab		
Extractor 1	1231	Raffelsberger 20	Community carri	Raffelsberger N;	https://dx.doi.org	https://cloud.lab		
Extractor 1	1281	Weil 2022	Mapping the em	Weil A.A.; Luiten	https://dx.doi.org	https://cloud.lab		

Main/Master "Overview Table" sorted by "Assignee".

- 45 Copy the **Main/Master "Overview Table"** for a data extractor over to the **"Overview Table"** in their respective **"Final_Data-Extraction_name"** worksheet.
- 46 OPTIONAL: Fill in the study number information across the columns in the **"Final_Data-Extraction_name" "Metadata Table"**. This can also be left to the data extractor using the form to fill in for themselves.
- 47 Do a **final review of the protocol** and polish up anything you can think of.
- 48 **REMINDER:** Make sure you copy over your solutions from **"Pilot_Data-Extraction_ANSWERS"** over to **"Pilot_Data-Extraction_you"** for the **"Overview Table"** and **"Metadata Table"**, and also upload the papers/supplementary files to the agreed upon storage location (e.g. Nextcloud).
- 49 Set-up permissions so data extractors have **"Editor"** access to their individual **"Final_Data-Extraction_name"** worksheet. You may also want to copy the **"Share"** link and send it to them directly using your normal means of communication so that they don't accidentally miss the notification.
- 50 **OPTIONAL:** Prepare a **"Final_Data-Extraction_RESULTS"** file that gives you an overview of everyone's progress. To do this you can make empty copies of the **"Overview Table"** and **"Metadata Table"** and then import the information from individual data extractor collection forms. If you have a large group you may want to keep separate **"Metadata Table"** tabs because having them all in one may be quite slow from the import function;

it'll be easier to combine these into one table in the end by copying the data values instead of importing.

Some benefits of doing this is you can go to a single location to get a sense of everyone's progress, and it updates "live" so if people end up going back to re-review a paper you won't have to re-extract their results.



Results from individual sheets being consolidated using "imports" into a final sheet.

Data Extraction (Final)

- 51 Locate your individual "**Final_Data-Extraction_name**" collection form and confirm you have edit permissions.
- 52 Review the "**README**" and the **data extraction protocol** for changes since the pilot.
- 53 Go to the "**Overview Table**" and find your papers using the URLs listed under "**Data Storage Location**".
- 54 Find all associated **supplementary data** using information within the paper(s) and/or the **DOI** listed in the "**Overview Table**". Look for data availability statements and reference to databases and database accessions. Download and label with the "**Study/Covidence #**" of the associated paper.
- 55 Upload the supplementary data to the associated "**Data Storage Location**".

**Note**

Consider doing this for all your papers before performing data extraction for the "Overview Table" and "Metadata Table". This way you can get into a rhythm with it and have this stage completely out of the way for the rest of your data extraction work.

- 56 Fill in the **"Overview Table"** for the study according to the guidance in your **protocol** and the **"Overview Legend"**.

Note

Depending on your preference, you can fill in the "Overview Table" for all studies and then do the "Metadata Table" for all studies, or you can fill in the "Overview Table" for one study and then immediately do the "Metadata Table" for the same study. We recommend you do the latter for the pilot, so that you keep your head in one "study" at a time, but when you get the hang of the process it tends to be more efficient to do the former.

- 57 Fill in the **"Metadata Table"** for the study according to the guidance in your **protocol** and the **"Metadata Legend"**.

Analysis

- 58 The purpose of this section is to give a general overview of the analyses performed with the first version of this protocol before the data visualizing stage. It is important to consider whether these outputs align with your designed data outcomes before you begin implementing the protocol so you have time and awareness to adjust the protocol as needed.

Note

These analyses were part of a scoping review that looked at the correlation between outbreak and/or surveillance research settings/objectives and different metadata types, as a precursor for more narrow focused systematic review(s) of contextual metadata in the future. For a systematic review and for evaluating significance more rigorous analysis is needed.

- 59 **Objective Coding**

Objectives are extracted out into their own table, distilled into themes/category codes which can then be used to organise the unstructured research objectives. The objective quotes can be left as raw data and are not needed for further analysis. The coded objectives can be added back into the Overview analysis table.

Covidence #	Objectives (separator = ;)	Assigned "Category" which we can generate from the final list of objectives/research questions/conclusions.
14	objective 1	Category 1
14	objective 242	Category 2
18	objective 50	Category 1

"OBJECTIVES" Intermediate Table.

60 Objectives Aggregated

- 60.1 Once the objective categories are defined, create a copy of your results "Overview Table" and then **classify each study or each study objective into the categories**. This can be done by adding a dropdown menu with the category options.

				Variables of Interest		
Covidence #	Study	Nextcloud (Data Storage Location)	Title	Type	Objectives (separator =)	Objectives Aggregated
3752	Alcolea-Medina	https://cloud.cidgo	The ongoing Strep Both		We, therefore, undertook a review of	Regional-level investiga... ▼
3761	Borges 2023	https://cloud.cidgo	Viral genetic clust	Outbreak Investig	Here, we combined epidemiological	National-level outbreak ... ▼
3763	Taouk 2023	https://cloud.cidgo	Intra- and interho	Surveillance	In this study, we examined the emerg	State, provincial, or terri... ▼
3766	Beale 2023	https://cloud.cidgo	Genomic epidem	Surveillance	We explored the value of WGS for su	National-level surveillan... ▼
3770	Mej  a-Ponce	https://cloud.cidgo	Genomic epidem	Surveillance	In this study, we expand upon this kn	National-level surveillan... ▼
3772	Kakooza 2023	https://cloud.cidgo	Genomic surveilla	Surveillance	The aims of the present study were t	International-level surve... ▼
3777	Baker 2023	https://cloud.cidgo	Use of High-Resol	Both	The objectives of our anal-ysis were t	Regional-level investiga... ▼
3785	Vestberg 2023	https://cloud.cidgo	Antimicrobial susc	Surveillance	The aim of this study was to investig	Regional-level surveilla... ▼
3786	Shen 2023	https://cloud.cidgo	A Nationwide Gen	Surveillance	Here, a comprehensive overview of t	National-level surveillan... ▼
3787	Ljungquist 2023	https://cloud.cidgo	Nationwide, popu	Surveillance	In this study we present essential mo	National-level surveillan... ▼
3788	Higgs 2023	https://cloud.cidgo	Population struct	Surveillance	In this study, we applied genome-sca	State, provincial, or terri... ▼
3789	Udaondo 2023	https://cloud.cidgo	Top-Down Genom	Surveillance	In this study, we applied a top-down	State, provincial, or terri... ▼
3796	Tsang 2023	https://cloud.cidgo	SARS-CoV-2 Outbr	Outbreak Investig	The aim of our study was to combine	Hospital-based outbrea... ▼
3800	Tessier 2023	https://cloud.cidgo	Genomic Epidem	Surveillance	The aim of this study was to describe	National-level surveillan... ▼
3803	deStMaurice 20	https://cloud.cidgo	Clinical, microbiol	Surveillance	Based on our findings, we continued	Regional-level surveilla... ▼

"Overview Table" objectives aggregated into categorical bins.

60.2 You can now **get summary counts of your results**, and use this information to look at the frequency of different contextual metadata types across your different objective categories.

<i>Objectives Aggregated</i>	<i>Value</i>
Community-level investigation	1
Community-level outbreak investigation	15
Community-level surveillance investigation	15
Hospital-based investigation	18
Hospital-based outbreak investigation	96
Hospital-based surveillance investigation	46
International-level investigation	1
International-level outbreak investigation	7
International-level surveillance investigation	10
National-level investigation	9
National-level outbreak investigation	29
National-level surveillance investigation	72
Regional-level investigation	11
Regional-level outbreak investigation	38
Regional-level surveillance investigation	41
State, provincial, or territorial-level investigation	2
State, provincial, or territorial-level outbreak investigation	11
State, provincial, or territorial-level surveillance investigation	12
Grand Total	434

Objective aggregate counts.



61 Metadata Table Value Preparation

- 61.1 **Check for empty cells and make sure none are “missing” data** (e.g. overlooked during data extraction as opposed to being “FALSE”).
- 61.2 **You’ll want to consolidate all the results into one table.** You can do this with import functions but it will be easier to copy/paste.
- 61.3 **Remove the “Definition” and “Example Subfield” columns, and clear all informative text from the “Required Subfield” columns.**
- 61.4 **Separate out the “Values” and “Motivations” into their own versions of the tables.** Continue these steps on the “Values” table.
- 61.5 **If a value is NOT “FALSE” (i.e. “TRUE” or “<value>”) then it is “TRUE”.**

Metadata Type (Field Name)	Required Subfields	#001	#002
Sample Collection			
Organism		FALSE	FALSE
purpose of sampling		TRUE	TRUE
geographical location (sample)		TRUE	TRUE
	geo_loc name (country) field	TRUE	TRUE
	geo_loc name (state/province/territory) field	TRUE	FALSE

Metadata Intermediate Table where values are TRUE/FALSE.

- 61.6 **Convert all “TRUE”s to 1 and all “FALSE”s to 0.**
- 61.7 **Replace empty “Required Subfield”s with N/A or some other null value or preference.**
- 61.8 **Repeat the “Metadata Type (Field Name)” value for the associated “Required Subfield”.**



Metadata Type (Field Name)	Required Subfields	#001	#002
Sample Collection			
Organism	N/A	0	0
purpose of sampling	N/A	1	1
geographical location (sample)	N/A	1	1
geographical location (sample)	geo_loc name (country) field	1	1
geographical location (sample)	geo_loc name (state/province/territory) field	1	0

Metadata Intermediate Table where values are 1/0 (binary).

- 61.9 **OPTIONAL: Pull out the "Category/Bin"/"Overarching Group" headers (if using hierarchical levels).** This is optional as they were added to facilitate data extraction/collection as opposed to being defined overarching categories of interest.
- 61.10 Now you have frequencies for the "Overarching Group", "Metadata Type (Field Name)", and "Required Subfields". The "Metadata Type (Field Name)" is covered by any rows that have "N/A" for required subfields because these were collected such that they are TRUE (1) if a required subfield is present, or independently if an example subfield was identified. So "N/A" rows can be "TRUE" (1) if at least one "Required Subfield" is "TRUE" (1) or even if all "Required Subfields" are "FALSE" (0). However, it should never be the case that an "N/A" row is "FALSE" (0) and one or more "Required Subfields" are "TRUE" (1), so we should perform a quick error check before proceeding.

ERROR CHECK: This can be done in Google Sheets using the formula

```
=if(and(or(<Required Subfield Range>),not(<N/A Range>)), "ERROR", "PASS")
```

A breakdown of the formula and a way to apply it within sheets is available [here](#).

- 61.11 **You'll need to add "Error Check" rows to apply the formula and then fix the "N/A" values that raise an "ERROR".** An example layout is available [here](#). You can easily copy/paste the formula across all the relevant columns and it'll update the range to be in the appropriate column.



Overarching Group	Metadata Type (Field Name)	Required Subfield	Values (pseudo data)		
			#001	#002	#003
ERROR CHECK			PASS	PASS	PASS
Sample Collection	isolate metadata	N/A	1	0	1
Sample Collection	isolate metadata	isolate ID	1	0	0
NO ERROR CHECK NECESSARY					
Sample Collection	organism	N/A	1	0	0
Sample Collection	purpose of sampling	N/A	1	0	1
ERROR CHECK			PASS	PASS	PASS
Sample Collection	geographical location (sample)	N/A	1	0	0
Sample Collection	geographical location (sample)	geo_loc name (country)	1	0	0
Sample Collection	geographical location (sample)	geo_loc name (state/province/territory)	1	0	0
ERROR CHECK			PASS	PASS	PASS
Sample Collection	sample collected by	N/A	1	0	1
Sample Collection	sample collected by	sample collected by	1	0	0
Sample Collection	sample collected by	sample collection date	1	0	0
Sample Collection	sample collected by	sample collection date precision	1	0	0
ERROR CHECK			PASS	PASS	ERROR
Sample Collection		N/A			
	sample identifiers		1	0	0
Sample Collection	sample identifiers	specimen collector sample ID	1	0	1
Sample Collection	sample identifiers	sample database accession ID	1	0	1

Error checking table for catching inconsistencies between categories/bins, fields, and subfields.

61.12 Remove the ERROR CHECK rows.

Overarching Group	Metadata Type (Field Name)	Required Subfield	Values (pseudo data)		
			#001	#002	#003
Sample Collection	isolate metadata	N/A	1	0	1
Sample Collection	isolate metadata	isolate ID	1	0	0
Sample Collection	organism	N/A	1	0	0
Sample Collection	purpose of sampling	N/A	1	0	1
Sample Collection	geographical location (sample)	N/A	1	0	0
Sample Collection	geographical location (sample)	geo_loc name (country)	1	0	0
Sample Collection	geographical location (sample)	geo_loc name (state/province/territory)	1	0	0
Sample Collection	sample collected by	N/A	1	0	1
Sample Collection	sample collected by	sample collected by	1	0	0
Sample Collection	sample collected by	sample collection date	1	0	0
Sample Collection	sample collected by	sample collection date precision	1	0	0
Sample Collection	sample identifiers	N/A	1	0	1
Sample Collection	sample identifiers	specimen collector sample ID	1	0	1
Sample Collection	sample identifiers	sample database accession ID	1	0	1
Sample Collection	sample collection methods	N/A	1	0	0
Sample Collection	sample collection methods	collection device	1	0	0
Sample Collection	sample collection methods	collection method	1	0	0
Sample Collection	sample collection environment	N/A	1	0	1
Sample Collection	sample collection environment	environmental site	1	0	1
Sample Collection	sample collection environment	environmental material	1	0	0
Sample Collection	sample collection environment	collection site host	1	0	1

"Metadata Table" results with error checking rows removed.

61.13 You can work with the table as is but it is probably easier to **separate this out into three tables** for the different levels.



1. "Required Subfield" Table (i.e. all "Required Subfield" rows except for the "N/A"s).
2. "Metadata Type (Field Name)" Table (i.e. all the "Required Subfield" "N/A"s).
3. OPTIONAL: "Overarching Group" Table (i.e. is "TRUE" (1) if any of the associated "Metadata Type (Field Name)" values are "TRUE" (1)).

Note

These header names can also be changed at any point, just try to be consistent with the current version of the "Metadata Table" to avoid miscommunication.

- 61.14 You can now easily **calculate sums for any of these tables to get the frequency counts across papers**. You can now also use the "Covidence #"s to cross reference the metadata table with the overview table. An example of which will be shown next in preparing a "Metadata x Objective" table.

Note

At this point it's probably best to shift to working in an R notebook, Python notebook, Tableau, or whatever you are most comfortable using for analysis.

62 Metadata x Objective Table Preparation (using R)

Goal: Indicate what metadata elements/categories are present per objective. We can also get frequencies of occurrence for different research objectives.

- 62.1 **Make a pivoted version of the metadata table that makes the columns your metadata element variables.**
- 62.2 **Cross reference with the overview table to bring in the "Objectives" column associated with the respective "Covidence #"**

Covidence #	Objectives	Metadata Element		
		isolate	organism	purpose of sampling
#001	<Objective 1> <Objective 2>	0	1	1
#002	<Objective 1> <Objective 2> <Objective 4>	0	1	1
#003	<Objective 2> <Objective 3>	0	1	1

"Metadata x Objective" intermediate table where objectives are grouped by study number.

62.3 **Now you want to separate "Objectives" out into their own rows.** All "Objectives" will have the same counts for the same "Covidence #". This can be done using **tidyr** ``separate_rows``.

Covidence #	Objectives	Metadata Element		
		isolate	organism	purpose of sampling
#001	<Objective 1>	0	1	1
#001	<Objective 2>	0	1	1
#002	<Objective 1>	0	1	1
#002	<Objective 2>	1	1	1
#002	<Objective 4>	1	1	1
#003	<Objective 2>	0	1	1

"Metadata x Objective" intermediate table where objectives are separated out into their own rows; one study may have several rows of objectives.

62.4 **You can get rid of or hide the "Covidence #" column.**

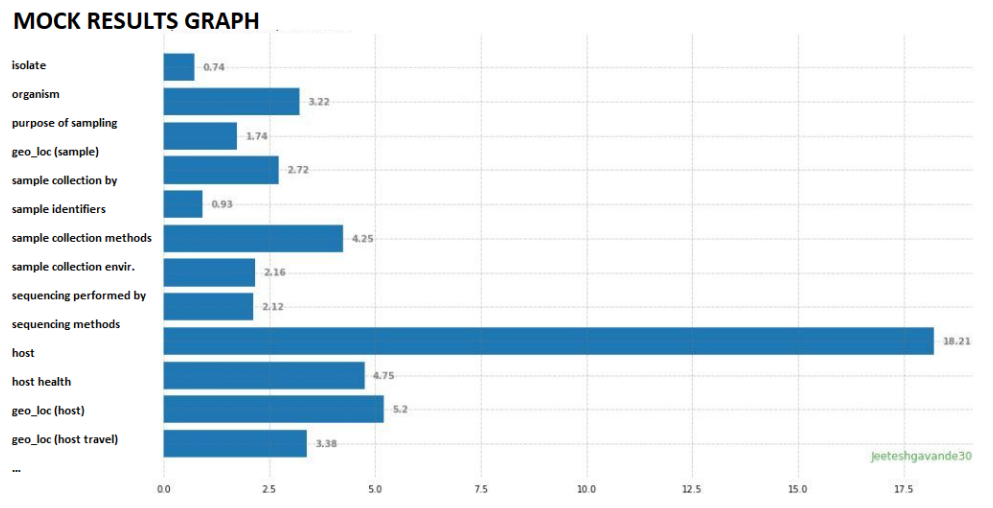
62.5 **Consolidate "Objectives" so there is only one row per objective code.** We recommend using **dplyr** ``group_by`` and ``summarise``.

Objectives	Metadata Element		
	isolate	organism	purpose of sampling
Category 1	2	3	3
Category 2	0	1	1
Category 3	0	2	2

"Metadata x Objective" final table.

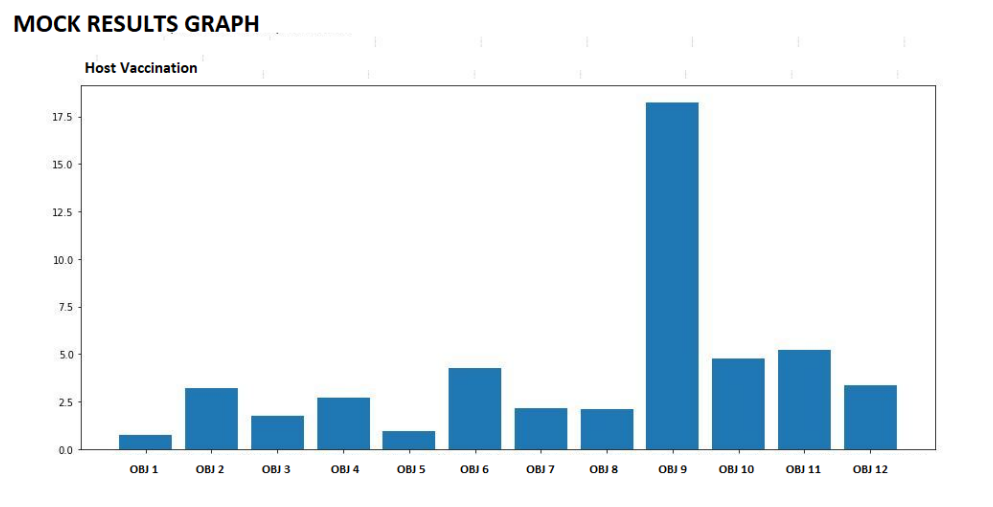
62.6 Results (Theoretical)

We can look at which objectives have the highest frequencies of which metadata.



Theoretical results for an objective category across different contextual metadata fields.

We can also look at which objectives are most frequently associated with a metadata category; may be more useful for less frequently used metadata items.



Theoretical results for how frequently different objectives occurred for a specific contextual metadata field (example: "Host Vaccination").

63 Motivations



We captured both positive and negative motivations. A positive motivation would be along the lines of "this metadata allowed us to do this", while a negative motivation could be "a limitation of that study was we couldn't study x because we didn't have y". It was determined that a "**Narrative Summary**" would be the appropriate approach; summarizing results using "evidence statements" (summaries of the evidence from the literature reviewed)¹. We also thought it would be interesting to do a "**Frequency Table**" to understand what contextual metadata was being most advocated for. Additionally or instead of the aforementioned analyses, quotes can simply be used as needed for discussion regarding the associated contextual metadata field(s).

Citation

(2022)

. LibGuides: Good review practice: a researcher guide to systematic review methodology in the sciences of food and health: 5. Data extraction.

https://ifis.libguides.com/systematic_reviews/data-extraction

LINK

64 Motivations - Narrative Summary

64.1 Extract all "Motivation" columns from the "Metadata & Motivation Extraction Table".

Metadata Type (Field Name)	#001	#002	#003
Sample Collection			
geographical location (sample)	"A key limitation of our study is that it did not include samples from all regions..."	FALSE	<example text>

"Motivation" Intermediate table showing raw data.

64.2 Copy and paste "transposed" into a new sheet.

	Metadata Type (Field Name)							
version 8	ID	Sample Coll	isolate meta	Isolate ID	organism	purpose of s	geographica	geo_loc nam
	14		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	18		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	30		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	55		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	63		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	75		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	77		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	78		FALSE	FALSE	"no studies have	"There is a pauc	FALSE	FALSE
	83		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	85		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	106		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	128		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	136		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	142		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	143		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	146		FALSE	FALSE	"Our investigatio	FALSE	FALSE	FALSE
	147		FALSE	FALSE	FALSE	"It is critical for h quickly identify ir the outbreak in c	FALSE	FALSE
	152		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

"Motivation" Intermediate table showing raw data transposed and with formatting removed.

64.3 Filter out "FALSE" data and pivot to a horizontal layout that makes it easy to read quotes and group by the associated contextual metadata field.

Metadata Field	Metadata Subfield	Study ID	Quote (From
sample collection	isolate metadata	424	The main strengths of this study are the population-based study des
sample collection	isolate metadata	599	The main limitations of our study were that isolates were submitted f
sample collection	isolate metadata	927	A limitation of this study is its dependence on archived isolates: as s
sample collection	Isolate ID	181	the availability of isolates from the circulating MSSA carriage popula
sample collection	Isolate ID	216	Strengths of this work include the inclusion of isolates from CA case:
sample collection	organism	78	no studies have reported on SARS-CoV2 genomic lineages/strains i
sample collection	organism	146	Our investigation highlights the importance of subtyping. WGS was e
sample collection	organism	3249	The virus spreads during the summer season, which is the active pe
sample collection	purpose of sampling	78	regions of Africa
sample collection	purpose of sampling	147	"It is critical for hospitals and health systems to not only quickly identify infected cases but also determine the source of the outbreak in order to mitigate the threat to patients and healthcare
sample collection	purpose of sampling	219	Genomic studies of gonococci are ideal for enhanced understanding
sample collection	purpose of sampling	256	limited by the absence of adequate (±50) and regular sampling at inc
sample collection	purpose of sampling	260	these studies suggest that the application of WGS within internatio
sample collection	purpose of sampling	323	Since the next-generation sequencing (NGS) technology has been a
sample collection	purpose of sampling	408	potential to improve the accuracy of cluster definitions and timeliness
sample collection	purpose of sampling	599	The main strength of our study is that it used data and iso-lates colle
sample collection	purpose of sampling	861	there was a dramatic increase in vanA VRE incidence in our institut
sample collection	geographical location (sample)	483	In the current study case, our data strongly suggest that WGS comb
sample collection	geographical location (sample)	593	Due to a lack of precise geolocation data (i.e. no postcode or GPS d

"Motivation" final table showing quotes, easily organized by contextual metadata fields or subfields.



64.4 **Qualitatively review the quotes and disqualify or relocate any that do not support the metadata field it was matched with.**

64.5 **Summarize the results into evidence statements.**

64.6 **Results (Theoretical)**

Discussion around the benefits of including a metadata element.

Example: Collecting clinical host metadata was key in the study's ability to evaluate risk factors in cases where breakthrough infection occurred.

Discussion around the limitations of not including a metadata element.

Example: One (or more) study claimed that the lack of clinical host metadata made it difficult to evaluate risk factors in cases where breakthrough infection occurred.

65 **Motivations - Frequency Counts**

Using the qualitatively reviewed "MOTIVATION Table" (under the Step for "Motivations - Narrative Summary"), converting any quotes into "TRUE" values and count the frequency in which they occur.

65.1 **Results (Theoretical)**

The frequency at which articles advocated for the inclusion of a metadata element.

EXAMPLE: 2 out of 3 articles advocated for the inclusion of the "geographical location" of sample collection.